Aalborg Universitet

AALBORG UNIVERSITY
DENMARK

Taking the Temperature of Sports Arenas

Automatic Analysis of People

Gade, Rikke

DOI (link to publication from Publisher):
10.5278/vbn.phd.engsci.00070

Publication date:
2014

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Gade, R. (2014). *Taking the Temperature of Sports Arenas: Automatic Analysis of People*. Aalborg Universitetsforlag. Ph.d.-serien for Det Teknisk-Naturvidenskabelige Fakultet, Aalborg Universitet https://doi.org/10.5278/vbn.phd.engsci.00070

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: May 06, 2023

TAKING THE TEMPERATURE OF SPORTS ARENAS

AUTOMATIC ANALYSIS OF PEOPLE

BY
RIKKE GADE

DISSERTATION SUBMITTED 2014

AALBORG UNIVERSITY
DENMARK

Taking the Temperature of Sports Arenas

Automatic Analysis of People

Ph.D. Dissertation

Rikke Gade

Aalborg University
Department of Architecture, Design and Media Technology
Rendsburggade 14
DK-9000 Aalborg

Thesis submitted: 28/11 2014

PhD supervisor: Professor Thomas B. Moeslund, Aalborg University

PhD committee: Associate Professor Claus Brøndgaard Madsen
Aalborg University (chairman)

Docent Henrik Karstoft
Aarhus University

Professor Graham Thomas
BBC Research & Development

PhD Series: Faculty of Engineering and Science, Aalborg University

ISSN: 2246-1248
ISBN: 978-87-7112-194-0

Published by:
Aalborg University Press
Skjernvej 4A, 2nd floor
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

© Copyright: Rikke Gade

Printed in Denmark by Rosendahls, 2014

Author CV

Rikke Gade received her M.Sc. in Engineering (Informatics with specialization in Vision, Graphics and Interactive Systems) from Aalborg University, Denmark in 2011. She worked as a research assistant at Aalborg University before starting her PhD study in December 2011 with the Visual Analysis of People Lab at the section of Media Technology, Aalborg University.

During the master and PhD studies she spent a semester abroad at University of Auckland, New Zealand, and she had a four month research stay at the Australian Centre for Visual Technologies, University of Adelaide, Australia.

She has been involved in supervision of undergraduate and graduate students in topics of image processing and computer vision. Her main research interest lies within computer vision for analysis of people.

Abstract

Measuring and mapping human activities are essential steps towards constructing an intelligent and efficient society. Using thermal imaging, the privacy issues often related to surveillance can be eliminated and public acceptance of such systems is easier to obtain.

The main focus of this thesis is automatic analysis of the use of sports arenas. This work is organised under three themes: Occupancy analysis, Activity recognition and Tracking. Finally, the thesis demonstrates how thermal imaging can also be applied efficiently for analysing humans in the Smart City.

The thermal camera is still considered a new sensor within the field of computer vision. This thesis starts by introducing the technology of the sensor and the different application areas. Two new methods for counting people are presented, the first method detecting objects in each frame independently. The second method exploits temporal information by estimating stable periods in the video and optimises the counting over a sequence of frames.

For activity recognition in sports arenas two different methods are presented. The first algorithm relies on positions of detected people. Heatmaps representing the occupancy of the arena are generated and classified between five different sports types using Fischer's Linear Discriminant. The second method for sports type classification is based on features extracted from short trajectories of each player. Four simple features; lifespan, total distance, distance span and mean speed are extracted and used for classification of five sports types.

Tracking of sports players is an important task in many applications, from recognition of activities to evaluation of performance. This thesis presents a real-time tracking algorithm based on Kalman filtering. It also introduces a method to improve tracking performance by constraining the number of trajectories produced by an offline tracking algorithm, and finally an algorithm using local updates to improve a global tracker.

At the end of this thesis five different applications of thermal imaging in the Smart City are presented. Methods for counting and tracking pedestrians are presented and applied, as well as a method for detecting potential near-collisions between cars and cyclists in large urban intersections.

Resumé

At kunne måle og kortlægge de menneskelige aktiviteter er et vigtigt skridt imod et intelligent og effektivt samfund. Med termiske kameraer kan bekymringerne om privatlivsrettighederne i forbindelse med overvågning undgås og det er dermed lettere at få befolkningens accept af et sådan system.

Det primære fokus for denne afhandling er automatisk analyse af brugen af sportshaller. Dette arbejde er organiseret under tre temaer: Analyse af belægning, Aktivitetsgenkendelse og Tracking. Endeligt demonstreres det også i denne afhandling hvordan termiske optagelser kan anvendes effektivt til at analysere menneskers færden i en "Smart City".

Det termiske kamera betragtes stadig som en ny sensor indenfor Computer Vision. Denne afhandling starter med at introducere teknologien bag disse sensorer og de forskellige anvendelsesområder. To nye metoder til at tælle mennesker præsenteres, hvor den første metode detekterer objekter i hver frame uafhængigt. Den anden metode udnytter temporal information ved at estimere stabile perioder i videoen og optimerer antallet over en sekvens af frames.

Til aktivitetsgenkendelse i sportshaller præsenteres to forskellige metoder. Den første algoritme beror på positionerne af de detekterede personer. Heatmaps, der repræsenterer den rumlige belægning i hallen, genereres og klassificeres mellem fem sportsgrene ved hjælp af Fischer's Linear Discriminant. Den anden metode for genkendelse af sportsgrene er baseret på features beregnet fra korte spor af hver spiller. Fire simple features; Levetid, samlet afstand, størst afstand og gennemsnitlig hastighed udtrækkes og bruges til klassificering af fem sportsgrene.

Tracking af spillere er en vigtig opgave i mange applikationer, fra genkendelse af aktiviteter til evaluering af præstationer. Denne afhandling præsenterer en realtidstrackingalgoritme baseret på Kalman filtrering. Der introduceres også en metode til at forbedre trackingresultaterne ved at indsnævre antallet af spor lavet af en offline trackingalgoritme og til sidst en algoritme der benytter lokale opdateringer til at forbedre en global tracker.

I slutningen af denne afhandling præsenteres fem forskellige anvendelser af termiske kameraer i en "Smart City". Metoder til at tælle og tracke fodgængere er præsenteret og anvendt, ligesom en metode til at detektere potentielle nærulykker mellem biler og cyklister i store vejkryds.

Contents

Thesis Details

Thesis Title:	Taking the Temperature of Sports Arenas - Automatic Analysis of People
Ph.D. Student:	Rikke Gade
Supervisor:	Prof. Thomas B. Moeslund, Aalborg University

The main body of this thesis consists of the following papers (the number refers to the chapter):

[2] Rikke Gade and Thomas B. Moeslund, "Thermal Cameras and Applications: A Survey," *Machine Vision and Applications*, vol. 25, no. 1, pp. 245–262, January 2014.

[4] Rikke Gade, Anders Jørgensen and Thomas B. Moeslund, "Occupancy Analysis of Sports Arenas Using Thermal Imaging," *Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP)*, pp. 277–283, February 2012.

[5] Rikke Gade, Anders Jørgensen and Thomas B. Moeslund, "Long-term Occupancy Analysis using Graph-based Optimisation in Thermal Imagery," *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3698–3705, June 2013.

[6] Rikke Gade and Thomas B. Moeslund, "Classification of Sports Types using Thermal Imagery," *Computer Vision in Sports*, Springer, January 2015.

[7] Rikke Gade and Thomas B. Moeslund, "Classification of Sports Types from Tracklets," *KDD workshop on Large-Scale Sports Analytics*, August 2014.

[8] Rikke Gade and Thomas B. Moeslund, "Thermal Tracking of Sports Players," *Sensors*, vol. 14, no. 8 pp. 13679–13691, July 2014.

[9] Rikke Gade, "Constrained Multi-Target Tracking for Thermal Imaging," Unpublished work in progress, November 2014.

[10] Anton Milan, Rikke Gade, Anthony Dick, Thomas B. Moeslund and Ian Reid, "Improving Global Multi-target Tracking with Local Updates," *ECCV workshop on Visual Surveillance and Re-Identification*, September 2014.

[11] Rikke Gade, Thomas B. Moeslund, Søren Zebitz Nielsen, Hans Skov-Petersen, Hans Jørgen Andersen, Kent Basselbjerg, Hans Thorhauge Dam, Ole B. Jensen, Anders Jørgensen, Harry Lahrmann, Tanja Kidholm Osmann Madsen, Esben Skouboe Bala and Bo Ø. Povey, "Thermal Imaging Systems for Real-Time Applications in Smart Cities," *International Journal of Computer Applications in Technology*, accepted for publication, October 2014.

[12] Esben Skouboe Poulsen, Hans Jørgen Andersen, Ole B. Jensen, Rikke Gade, Tobias Thyrrestrup and Thomas B. Moeslund, "Controlling Urban Lighting by Human Motion Patterns - results from a full scale experiment," *Proceedings of the ACM International Conference on Multimedia*, pp. 339–347, October 2012.

[13] Søren Zebitz Nielsen, Rikke Gade, Thomas B. Moeslund and Hans Skov-Petersen, "Taking the temperature of Pedestrian Movement in Public Spaces," *Transportation Research Procedia - Conference on Pedestrian and Evacuation Dynamics*, vol. 2, pp. 660–668, October 2014.

In addition to the main papers, the following publications have also been made.

[A] Rikke Gade, Cecilie Breinholm Christensen, Rasmus Krogh Jensen, Thomas B. Moeslund, Henrik Harder, "Analyse af Adfærd i Idrætsfaciliteter," *Forum for Idræt*, vol. 29, no. 1, pp. 121–133, December 2013.

[B] Esben Skouboe Poulsen, Hans Jørgen Andersen, Rikke Gade, Ole B. Jensen and Thomas B. Moeslund, "Using Human Motion Intensity as Input for Urban Design," *Constructing Ambient Intelligence - Communications in Computer and Information Science*, vol. 277, pp. 128–136, November 2012.

[C] Rikke Gade, Anders Jørgensen, Thomas B. Moeslund and Rasmus Krogh Jensen, "Automatic Analysis of Sports Arenas," *EASM conference*, September 2012.

[D] Rikke Gade and Thomas B. Moeslund, "Sports Type Classification using Signature Heatmaps," *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 999–1004, June 2013.

This thesis has been submitted for assessment in partial fulfillment of the PhD degree. The thesis is based on the submitted or published scientific papers which are listed above. Parts of the papers are used directly or indirectly in

the extended summary of the thesis. As part of the assessment, co-author statements have been made available to the assessment committee and are also available at the Faculty. The thesis is not in its present form acceptable for open publication but only in limited and closed circulation as copyright may not be ensured.

Preface

This thesis is submitted as a collection of papers in partial fulfillment of a PhD study at the Section of Media Technology, Department of Architecture, Design and Media Technology, Aalborg University, Denmark. The thesis is organised in six parts. The first part contains the motivation, background and a summary of the included papers. Part 2-5 organise the papers in four themes: Occupancy analysis, Activity recognition, Tracking sports players and Smart City applications. The final part concludes the thesis.

The work has been conducted from December 2011 to November 2014 as part of the project *"Bedre Brug af Hallen"* funded by *Nordea-fonden* and *Lokale og Anlægsfonden*. The project collaborated with *Aalborg Kommune*. I am very grateful for their support and help in relation to get access to sports arenas for capturing the large amount of video data we have used.

I will like to thank everyone at the Section of Media Technology for their support, technical discussions and small chats. Special thanks goes to my supervisor Thomas Moeslund, for encouraging me to undertake this project and for his positive, friendly and highly qualified supervision. Thanks to my colleague Anders Jørgensen for collaborating on this research project and for his patience during numerous hours of data capturing. I also wish to thank my colleagues during my stay at the Australian Centre for Visual Technologies, University of Adelaide, for warmly welcoming me and for their collaboration and support during my stay.

Rikke Gade
Aalborg University, November 28, 2014

Part I

Introduction

Chapter 1

Introduction

With a society constantly looking for ways to save costs, improve efficiency, and raise the standard of living, automation is one of the popular means applied. Robots take over trivial assembly tasks at factories, and machines will greet you at the supermarket. In the same way, the monotonously job of monitoring surveillance videos is being replaced with automatic computer vision algorithms.

The notion of surveillance is generally being related to controlling and monitoring people often with the purpose of preventing or investigating crime. However, surveillance can also be used for acquisition of large amounts of anonymous data, which can be applied for evaluating the general behaviour and activities in relation to the surrounding space or building.

Traditionally, the analysis of use of both indoor and outdoor facilities has been conducted by manual observations and written notes. Being a tedious and expensive job, automation of this process is important in order to get more data and objective measurements for a lower cost. Compared to other applied technologies, such as RFID and GPS, cameras normally have higher spatial resolution and are non-intrusive to the users of the facilities. This also means that it is not necessary to sample the user group, all individuals can be analysed with no extra cost.

As the technology of vision sensors has evolved, the computer vision research is also starting to take advantage of the new image modalities available. Various types of depth sensors, as well as near-infrared night vision sensors, have been applied in many applications, for both research, commercial and industrial purposes. The thermal camera, sensing the long-waved infrared spectrum, were originally developed for military use as a night vision instrument. Since

commercialised in late 1980s they have been used for temperature measuring of buildings and components in industry, but slowly, the use of thermal cameras in typical surveillance systems has increased. The independence of light is encouraging the use of thermal cameras in outdoor environments and in any application with need for robust performance day and night.

Most important here, the privacy preserving nature of the thermal sensor makes it suited for analysis of humans in sensitive applications. Throughout this thesis the focus is applications in public sports arenas. With users from young children to elderly people, privacy is an important issue. Therefore, thermal cameras will be applied, to preclude identification of individual people.

The remaining part of this introduction will firstly specify the focus of the thesis. The following chapter consists of the published journal paper "Thermal Cameras and Applications: A Survey". This chapter will provide a thorough review of the technology of thermal cameras, as well as a literature survey of the applications of thermal imaging. Finally, chapter 3 will provide a summary of all included papers in this thesis and sum up the contributions made to the field of computer vision.

1.1 Focus of this thesis

The work presented in this thesis deals with automatic detection and analysis of people observed in thermal video. The research has been conducted under three main themes: *Occupancy analysis, Activity recognition* and *Tracking sports players*.

Detecting people is the first step in every analysis of humans. For applications in sports arenas with various natural sports activities observed, the methods must be robust to any pose change and heavy occlusions between people. The occupancy analysis deals with these challenges. After occupancy, the activities performed in a sports arena are analysed. In the third part a very popular research area is dealt with; tracking of humans. The focus of this part is narrowed down to methods applicable to thermal imaging and sports players.

The work presented in the first three parts of this thesis is directly related to the research project "Better use of sports arenas" funded by *Nordea-fonden* and *Lokale- og Anlægsfonden*. The aim of this project was to investigate, develop and apply new methods for analysing the use of sports arenas. The research presented in this thesis represents the scientific content of the project, but is has been closely coupled to the practical aspects. A very positive outcome of this relation has been the access to several public sports arenas, from which all thermal sports data is captured. Thus, all data used in this research is captured from real everyday activities with regular users of the facilities.

The practical applicability of the methods presented in this work is reflected in part V. We present here a few applications where we in collaboration

with both internal and external partners have demonstrated how detection and tracking of humans can be applied in the Smart City.

Chapter 2

Thermal Cameras and Applications: A Survey

Rikke Gade and Thomas B. Moeslund

The paper has been published in
Machine Vision and Applications Vol. 25(1), pp. 245–262, January 2014.

© 2014 Springer
The layout has been revised.

Abstract

Thermal cameras are passive sensors that capture the infrared radiation emitted by all objects with a temperature above absolute zero. This type of camera was originally developed as a surveillance and night vision tool for the military, but recently the price has dropped, significantly opening up a broader field of applications. Deploying this type of sensor in vision systems eliminates the illumination problems of normal greyscale and RGB cameras.

This survey provides an overview of the current applications of thermal cameras. Applications include animals, agriculture, buildings, gas detection, industrial, and military applications, as well as detection, tracking, and recognition of humans. Moreover, this survey describes the nature of thermal radiation and the technology of thermal cameras.

2.1 Introduction

During the last couple of decades, research and development in automatic vision systems has been rapidly growing. Visual cameras, capturing visible light in greyscale or RGB images, have been the standard imaging device. There are, however, some disadvantages to use these cameras. The colours and visibility of the objects depend on an energy source, such as the sun or artificial lighting. The main challenges are therefore that the images depend on the illumination, with changing intensity, colour balance, direction, etc. Furthermore, nothing can be captured in total darkness. To overcome some of these limitations and add further information to the image of the scene, other sensors have been introduced in vision systems. These sensors include 3D sensors [1–3] and near infrared sensors [4]. Some of the devices are active scanners that emit radiation, and detect the reflection of the radiation from an object. Night vision devices, for example, use active infrared cameras, which illuminate the scene with near infrared radiation (0.7–1.4 μm) and capture the radiation of both the visible and the near infrared electromagnetic spectrum. Such active sensors are less dependent on the illumination. Stereo vision cameras are passive 3D sensors, but as they consist of visual cameras, they also depend on the illumination.

The described sensors indicate that some of the disadvantages of visual cameras can be eliminated by using active sensoring. However, in many applications, a passive sensor is preferred. In the mid- and long-wavelength infrared spectrum (3–14 μm), radiation is emitted by the objects themselves, with a dominating wavelength and intensity depending on the temperature. Thereby they do not depend on any external energy source. Thermal cameras utilise this property and measure the radiation in parts of this spectrum. Figure 2.1 shows an example of the same scene captured with both a visual and a thermal camera. The thermal image is shown as a greyscale image, with bright pixels for hot objects. The humans are much easier to distinguish in the thermal image, while the colours and inanimate objects, like chairs and tables, are invisible.

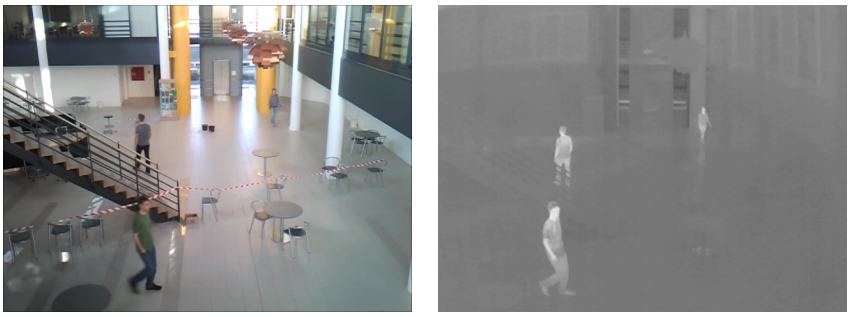


Fig. 2.1: Visible and thermal image of the same scene.

A special detector technology is required to capture thermal infrared radiation. Originally it was developed for night vision purposes for the military, and the devices were very expensive. The technology was later commercialised and has developed quickly over the last few decades, resulting in both better and cheaper cameras. This has opened a broader market, and the technology is now being introduced to a wide range of different applications, such as building inspection, gas detection, industrial appliances, medical science, veterinary medicine, agriculture, fire detection, and surveillance. This wide span of applications in many different scientific fields makes it hard to get an overview. This paper aims at providing exactly such an overview and in addition provides an overview of the physics behind the technology.

The remaining part of this survey consists of the following sections: Section 2.2 describes the physics of thermal radiation and Section 2.3 explains the technology of the cameras. Description of the application areas and a survey of the work done in the different areas are found in Section 2.4. In Section 2.5 it is discussed how to fuse the thermal images with other image modalities, and the application areas for fused systems are surveyed. Finally, Section 2.6 summarizes and discusses the use of thermal cameras.

2.2 Thermal Radiation

Infrared radiation is emitted by all objects with a temperature above absolute zero. This is often referred to as thermal radiation. This section will go through the source and characteristics of this type of radiation.

2.2.1 Electromagnetic Spectrum

Infrared radiation lies between visible light and microwaves within the wavelength spectrum of 0.7–1000 μm as illustrated in Figure 2.2. The infrared spectrum can be divided into several spectral regions. There exist different subdivision schemes in different scientific fields, but a common scheme is shown

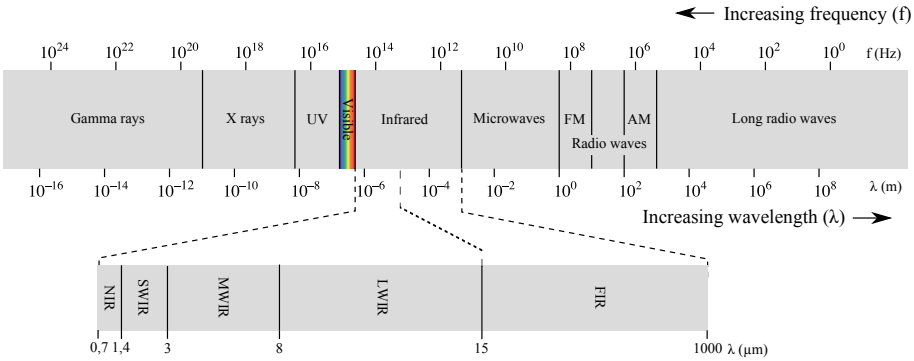


Fig. 2.2: The electromagnetic spectrum with sub-divided infrared spectrum.

Division Name	Abbreviation	Wavelength
Near-infrared	NIR	0.7–1.4 μm
Short-wavelength infrared	SWIR	1.4–3 μm
Mid-wavelength infrared	MWIR	3–8 μm
Long-wavelength infrared	LWIR	8–15 μm
Far-infrared	FIR	15–1000 μm

Table 2.1: Infrared sub-division.

in Table 2.1 [5]. The mid-wavelength and long-wavelength infrared are often referred to as thermal infrared (TIR) since objects in the temperature range from approximately 190 K to 1000 K emit radiation in this spectral range.

The atmosphere only transmit radiation with certain wavelengths, due to the absorption of other wavelengths in the molecules of the atmosphere. CO_2 and H_2O are responsible for most of the absorption of infrared radiation [6]. Figure 2.3 illustrates the percentage of transmitted radiation depending on the wavelength, and states the molecule that is responsible for the large transmission gaps.

Due to the large atmospheric transmission gap between 5–8 μm, there is no reason for cameras to be sensitive in this band. The same goes for radiation above 14 μm. A typical spectral range division for near-infrared and thermal cameras is shown in Table 2.2.

Division Name	Abbreviation	Wavelength
Short-wave	SWIR	0.7–1.4 μm
Mid-wave	MWIR	3–5 μm
Long-wave	LWIR	8–14 μm

Table 2.2: Infrared sub-division for cameras.

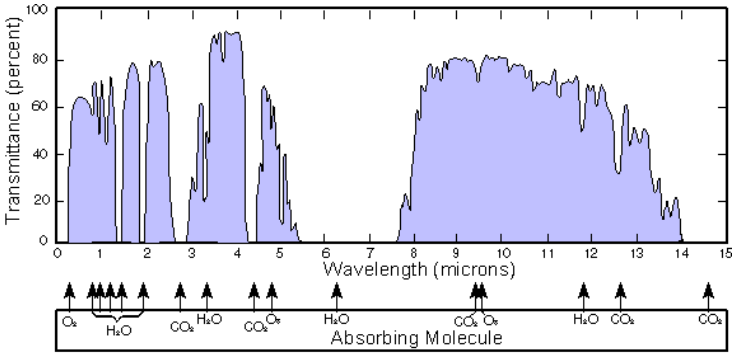


Fig. 2.3: Atmospheric transmittance in part of the infrared region [7].

2.2.2 Emission and Absorption of Infrared Radiation

The radiation caused by the temperature T of an object is described by Planck's wavelength distribution function [8]:

$$I(\lambda, T) = \frac{2\pi hc^2}{\lambda^5 \left(e^{hc/\lambda k_B T} - 1\right)}, \tag{2.1}$$

where λ is the wavelength, h is Planck's constant $(6.626 \times 10^{-34} Js)$, c the speed of light $(299,792,458 m/s)$ and k_B Boltzmann's constant $(1.3806503 \times 10^{-23} J/K)$.

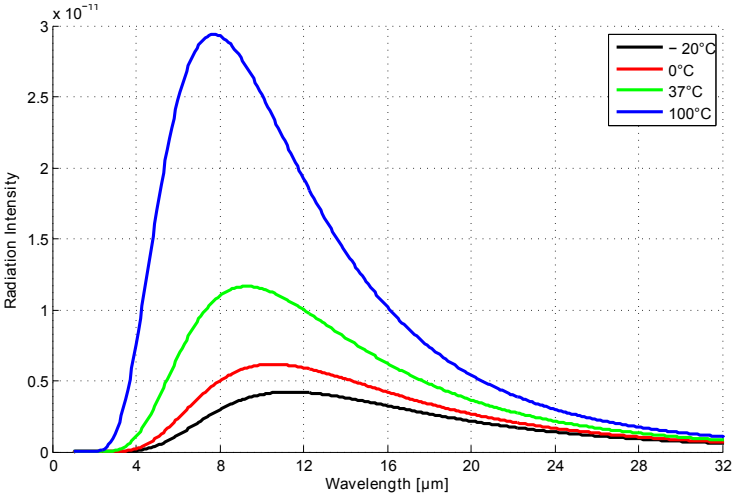


Fig. 2.4: Intensity of black body radiation versus wavelength at four temperatures.

As can be seen in Figure 2.4, the intensity peak shifts to shorter wavelengths as the temperature increases, and the intensity increases with the temperature.

For extremely hot objects the radiation extends into the visible spectrum, e.g., as seen for a red-hot iron. The wavelength of the intensity peak is described by Wien's displacement law [8]:

$$\lambda_{max} = \frac{2.898 \times 10^{-3}}{T}. \tag{2.2}$$

Planck's wavelength distribution function, Equation 2.1, describes the radiation from a black body. Most materials studied in practical applications are assumed to be so called grey bodies, which have a constant scale factor of the radiation between 0 and 1. This factor is called the emissivity. For instance, polished silver has a very low emissivity (0.02) while human skin has an emissivity very close to 1 [9]. Other materials, such as gases, are selective emitters, which have specific absorption and emission bands in the thermal infrared spectrum [6]. The specific absorption and emission bands are due to the nature of the radiation, as described in the next section.

2.2.3 Energy States of a Molecule

The energy of a molecule can be expressed as a sum of four contributions [8]: electronic energy, due to the interactions between the molecule's electrons and nuclei; translational energy, due to the motion of the molecule's centre of mass through space; rotational energy, due to the rotation of the molecule about its centre of mass; and vibrational energy, due to the vibration of the molecule's constituent atoms:

$$E = E_{el} + E_{vib} + E_{rot} + E_{trans}. \tag{2.3}$$

The translational, rotational, and vibrational energies contribute to the temperature of an object.

The possible energies of a molecule are quantized, and a molecule can only exist in certain discrete energy levels. Figure 2.5 illustrates the relation between the electronic, vibrational, and rotational energy levels. The contribution from the translational energy is very small and is not included in this illustration.

Electromagnetic radiation can be absorbed and emitted by molecules. Incident radiation causes the molecule to rise to an excited energy state, and when it falls back to the ground state, a photon is released. Only photons with specific energies, equal to the difference between two energy levels, can be absorbed. Visible light usually causes electron transitions, with rising or falling electronic energy level. Just as for visible light, infrared light can cause transitions in the vibrational or rotational energy levels. All objects emit infrared radiation corresponding to their temperature. If more radiation is absorbed than emitted, the temperature of the molecule will rise until equilibrium is re-established. Likewise, the temperature will fall if more radiation is emitted than absorbed, until equilibrium is re-established.

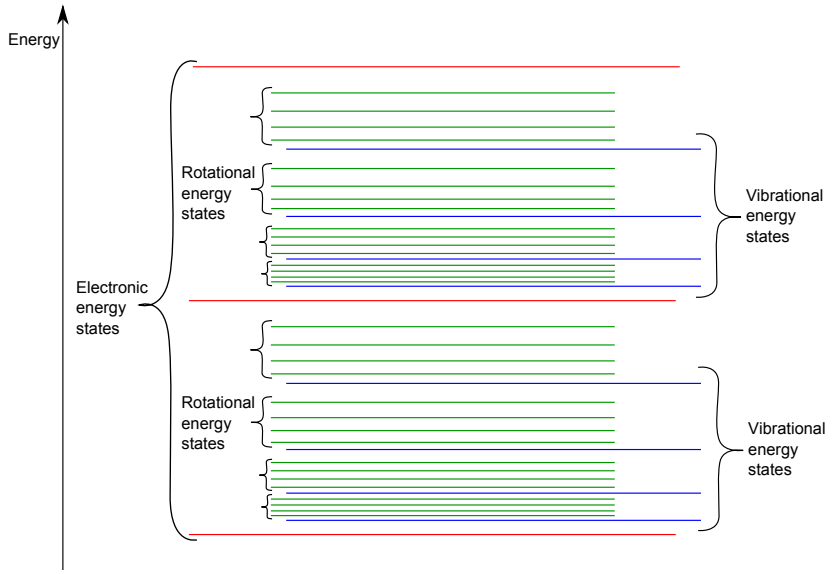


Fig. 2.5: Simplified illustration of the electronic, vibrational, and rotational energy states. Each line illustrates a discrete energy level that the molecule can exist in.

2.3 Thermal Cameras

Although infrared light was discovered by William Herschel around 1800, the first infrared scanning devices and imaging instruments were not build before the late 1940s and 1950s [10]. They were built for the American military for the purpose of night vision. The first commercial products were produced in 1983 and opened up a large area of new applications.

The measurement instruments available today can be divided into three types: point sensors, line scanners, and cameras.

2.3.1 Camera Types

Infrared cameras can be made either as scanning devices, capturing only one point or one row of an image at a time, or using a staring array, as a two-dimensional infrared focal plane array (IRFPA) where all image elements are captured at the same time with each detector element in the array. Today IRFPA is the clearly dominant technology, as it has no moving parts, is faster, and has better spatial resolution than scanning devices [10]. Only this technology is described in the following.

The detectors used in thermal cameras are generally of two types: photon detectors or thermal detectors. Photon detectors convert the absorbed electromagnetic radiation directly into a change of the electronic energy distribution in a semiconductor by the change of the free charge carrier concentration.

Thermal detectors convert the absorbed electromagnetic radiation into thermal energy causing a rise in the detector temperature. Then the electrical output of the thermal sensor is produced by a corresponding change in some physical property of material, e.g., the temperature-dependent electrical resistance in a bolometer [6].

The photon detector typically works in the MWIR band where the thermal contrast is high, making it very sensitive to small differences in the scene temperature. Also with the current technology the photon detector allows for a higher frame rate than thermal detectors. The main drawback of this type of detector is its need for cooling. The photon detector needs to be cooled to a temperature below 77 K in order to reduce thermal noise. This cooling used to be done with liquid nitrogen, but now is often implemented with a cryocooler. There is a need for service and replacement for the cryocooler due to its moving parts and helium gas seals. The overall price for a photon detector system is therefore higher than a thermal detector system, both its initial costs and its maintenance.

A thermal detector measures radiation in the LWIR band and can use different detector types, which will be described in the next section.

Thermal Detector Types

Uncooled thermal detectors have been developed mainly with two different types of detectors: ferroelectric detectors and microbolometers. Ferroelectric detectors take advantages of the ferroelectric phase transition in certain dielectric materials. At and near this phase transition, small fluctuations in temperature cause large changes in electrical polarization [11]. Barium Strontium Titanate (BST) is normally used as the detector material in the ferroelectric detectors.

A microbolometer is a specific type of resistor. The materials most often used in microbolometers are Vanadium Oxide (VOx) and Amorphous silicon (a-Si). The infrared radiation changes the electrical resistance of the material, which can be converted to electrical signals and processed into an image.

Today it is clear that microbolometers have more advantages over the ferroelectric sensors and the VOx technology has gained the largest market share. First of all, microbolometers have a higher sensitivity. The noise equivalent temperature difference (NETD), specifying the minimum detectable temperature difference, is 0.039 K for VOx compared to 0.1 K for BST detectors [11]. Microbolometers also have a smaller pixel size on the detector, allowing a higher spatial resolution. Furthermore, BST detectors suffer from a halo effect, which can often be seen as a dark ring around a bright object, falsely indicating a lower temperature [11]. An example of the halo effect is shown in Figure 2.6.

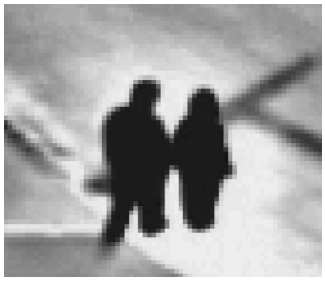


Fig. 2.6: Thermal image showing bright halo around a dark person [12].

2.3.2 The Lens

Since glass has a very low transmittance percentage for thermal radiation, a different material must be used for the lenses. Germanium is used most often. This is a grey-white metalloid material which is nearly transparent to infrared light and reflective to visible light. Germanium has a relatively high price, making the size of the lens important.

The f-number of an optical system is the ratio of the lens's focal length to the diameter of the entrance pupil. This indicates that a higher f-number reduces the price of the lens, but at the same time, when the diameter of the lens is reduced, a smaller amount of radiation reaches the detector. In order to maintain an acceptable sensitivity, uncooled cameras must have a low f-number. For cooled cameras, a higher f-number can be accepted, because the exposure time can be increased in order to keep the same radiation throughput. These properties of the lens cause the price for uncooled cameras to increase significantly with the focal length, while the price for cooled cameras only increases slightly with the focal length. For very large focal lengths, cooled cameras will become cheaper than uncooled cameras [13].

2.3.3 Camera Output

Modern thermal cameras appear just like visual video cameras in terms of shape and size. Figure 2.7 shows an example of a thermal network camera.

Fig. 2.7: Example of an uncooled thermal camera, AXIS Q1921.

The data transmission typically takes place via USB, Ethernet, FireWire, or RS-232. The images are represented as greyscale images with a depth from 8 to 16 bit per pixel. They are, however, often visualised in pseudo colours for better visibility for humans. Images can be compressed with standard JPEG and video can be compressed with H264 or MPEG [14]. Analogue devices use the NTSC or PAL standards [15]. Some handheld cameras are battery-driven, while most of the larger cameras need an external power supply or Power over Ethernet.

The thermal sensitivity is down to 40 mK for uncooled cameras and 20 mK for cooled devices. The spatial resolution of commercial products varies from 160×120 pixels to 1280×1024 pixels, and the field of view varies from $1°$ to $58°$ [16–19].

2.4 Application Areas

The ability to 'see' the temperature in a scene can be a great advantage in many applications. The temperature can be important to detect specific objects, or it can provide information about, e.g., type, health, or material of the object. This section will survey the applications of thermal imaging systems with three different categories of subjects: animals and agriculture, inanimate objects, and humans.

2.4.1 Animals and Agriculture

Animals

Warm-blooded animals, such as humans, try to maintain a constant body temperature, while cold-blooded animals adapt their temperature to their surroundings. This property of warm-blooded animals makes them stand out from their surroundings in thermal images. Warm-blooded animals can warm their body by converting food to energy. To cool down, they can sweat or pant to lose heat by water evaporation. The radiation of heat from animals depends on their insulation, such as hair, fur, or feathers, for example. The temperature distribution over the body surface can be uneven, depending on blood-circulation and respiration. In the studies of wild animals thermal imaging can be useful for diagnosis of diseases and thermoregulation, control of reproductive processes, analysis of behaviour, as well as detection and estimation of population size [20].

Diseases will often affect the general body temperature, while injuries will be visible at specific spots, e.g., caused by inflammations. Thermal imaging has thereby been proven to work as a diagnosis tool for some diseases of animals. In [21] it was observed that the temperature in the gluteal region of dairy cattle increases when the animal becomes ill and this could be detected in thermal images prior to clinical detection of the disease. If the observed animals are wild, the method of examining for a disease should be without contact with

the animals. In [22] thermal cameras are used for detecting sarcoptic mange in Spanish ibex. Although conventional binoculars have higher sensitivity over a greater distance, thermal cameras can give indication of the prevalence of the disease in a herd. Thermal imaging could also be used to detect diseases among other wild animals, in [23] it is found that rabies can be detected in raccoons by observing the temperature of the nose.

The stress level of animals before slaughtering is important to the meat quality. The stress level is correlated with the blood and body temperature of the animal. It is therefore important to monitor and react to a rising temperature, e.g., during transport. The work of [24] measures the temperature of pigs' ears and finds that it is positively correlated with the concentration of cortisol and the activity of creatine kinase.

Thermal imaging can be beneficial when diagnosing lameness in horses. [25] suggests using thermal imaging for detecting inflammations and other irregularities, especially in the legs and hoofs of horses. Figure 2.8 shows an example of inflammation in the leg.

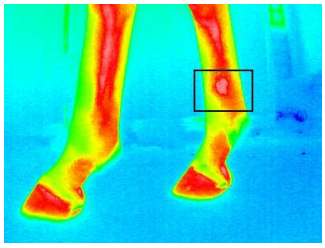


Fig. 2.8: The thermal image reveals inflammation in the leg of a horse. The inflamed area is marked with a black box.

Analysis of the thermodynamic characteristics in ectotherm animals, such as frogs, has been carried out in [26]. They measure the temperature of different body parts of frogs during heating from $8°C$ (artificial hibernation) to $23°C$ (artificial arousal). In such experiments it is a great advantage that the measurements are taken without harming or touching the animal.

Large animals can pose a risk for traffic if they run onto the road. They can often be hard to spot with the eye, specially in the dark or haze, also if they are camouflaged beside the road. Deer are some of the animals that can be a threat to safety on the roads. In [27], they propose a system for detecting and tracking deer from a vehicle, in order to avoid collisions. Some car brands have implemented thermal cameras and screens in their cars for manual detection of unexpected hot objects [28].

Wild animals have a high risk of being injured or killed during farming routines with modern high-efficiency farming equipment. Therefore [29] proposes automatic analysis of thermal images for detection of animals hidden in the vegetation. They use a pre-processing step by filtering the image with the Laplacian of Gaussian, before using adaptive thresholding for detecting the

animal.

Agriculture and Food

Thermal imaging systems have various applications in the agriculture and food industry. They are suitable in the food industry due to their portability, realtime imaging, and non-invasive and non-contact temperature measurement capability [30]. In food quality measurement, it is important to use a non-destructive method to avoid waste.

The two papers [31] and [30] review the use of thermal imaging in the agriculture and food industry, including both passive thermography (measuring the temperature of the scene) and active thermography (adding thermal energy to an object, and then measuring the temperature). Passive thermography is mostly used for temperature control in food manufacturing and for monitoring heat processes. Active thermography of food objects can give information about the quality, such as damage and bruises in fruits and vegetables. Bruises can be detected using active thermal imaging, due to the different thermodynamic properties in sound and bruised tissue. Thermal imaging has been applied in [32] to detect fungal infections in stored wheat. It could discriminate between healthy and infected wheat, but not between different fungal species. In [33], they classify healthy and fungal infected pistachio kernels.

2.4.2 Inanimate Objects

Inanimate objects do not maintain a constant temperature. Their temperature depends on both the surrounding temperature, and the amount of added energy that generates heat. Thermal images of inanimate objects depict the surface temperature of the scene. But even in a scene in equilibrium, differences in the image can be observed due to different emissivities of the observed surfaces. Thus thermal imaging can be used for analysing both temperature and material.

Building Inspection

Thermal cameras have been used for years for inspecting heat loss from buildings, and special hand-held imaging devices have been developed with this application in mind. Figure 2.9 shows an example of a thermal image of a building.

Normally the inspection of buildings requires manual operation of the camera and interpretation of the images to detect heat loss, e.g., as described in [35]. More automatic methods are also being investigated. In [36], an Unmanned Aerial Vehicle (UAV) is used for inspection of buildings, and the system automatically detects the heat loss from windows. Another system has been proposed, which automatically maps the images to a 3D model, eliminates windows and doors, and detects regions with high heat loss on the facade [37–39]. A thermal system has also been proposed for detecting roof leakage [40].

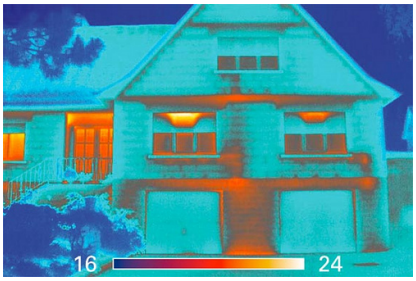


Fig. 2.9: Thermal image of a building, showing a higher amount of heat loss around windows and doors [34].

Besides the detection of heat loss, thermal imaging has also been used to detect other problems behind the surface: [41] proves that thermal imaging can be used to detect debonded ceramic tiles on a building finish. Termites can also be found by inspection with a thermal camera, as they produce unusual heat behind the surface in buildings [42].

For some ancient buildings, it is of interest to monitor the wall's hidden structure, the surface status, and moisture content, which can be done with a thermal camera [43]. The documentation of a building's status can also be done by combining visual and thermal images [44].

Another interesting application related to buildings is the one presented in the book *Mobile Robot Navigation with Intelligent Infrared Image* [45]. They present an outdoor robot system equipped with a thermal camera and an ultrasound sensor. In order to move around safely, the robot should be able to classify typical outdoor objects, such as trees, poles, fences, and walls, and make decisions about how to go around them. The classification of these non heat-generating objects is based on their physical properties, such as emissivity, that influence their thermal profile.

Gas Detection

Gasses are selective emitters, which have specific absorption and emission bands in the infrared spectrum, depending on their molecular composition. By using instruments able to measure selectable narrow infrared bands, it is possible to measure the radiation in the absorption band of a specific gas. As the radiation is absorbed by the gas, the observed area would appear as a cool cloud (usually dark) if the gas is present.

Using optical bandpass filters is applied for measuring carbon monoxide in [46]. Using a thermal spectrometer, a number of bands can be measured concurrently to analyse the gas content in the scene. In [47], they use 12 spectral bands distributed from 8.13μm to 11.56μm to detect an anomalous gas and track it in the image to locate the source of the gas leak. [48] tests a method for detecting gas leakage in landfills based on the temperature measurements of

a thermal camera (8–13 μm). They conclude that it is possible, but depends on the weather conditions and climate. [49] detects gas leaks of ammonia, ethylene, and methane by measuring the spectral region 7–13 μm. Volcanic ash particles can also be detected by measuring five spectral bands between 7–14 μm [50].

Industrial Applications

In most electrical systems, a stable temperature over time is important in order to avoid system break-downs. Sudden hot spots can indicate faulty areas and connections, e.g., in electric circuits and heating systems. It would obviously be of great value if devices that are starting to over-heat could be detected before they break down. One of the reasons for using thermal imaging for temperature measurement is that it is not in contact with the target. Thermal imaging can be applied as a diagnosis tool for electrical joints in power transmission systems [51], and for automatic detection of the thermal conditions of other electrical installations [52]. It can also be used to evaluate specific properties in different materials. In [53], the erosion resistance of silicon rubber composites is evaluated using a thermal camera. In metal sheet stamping processes, the mechanical energy is converted into thermal energy. An unexpected thermal distribution can be an indication of malfunctions in the object. Therefore [54] proposes a system that compares the thermal images to a simulated thermal pattern in order to find a diagnosis for the object. For more complicated objects, a 3D model is generated. [55] uses thermal imaging for measuring the molten steel level in continuous casting tundish.

For race cars, tire management is extremely important, and one of the main parameters of a tire is its temperature. [56] proposes the use of a thermal camera for dynamic analysis of the temperature of the tires during a race.

Fire Detection and Military

Automatic systems for detecting objects or situations that could pose a risk can be of great value in many applications. A fire detection system can be used for mobile robots. [57] proposes a system using a pan-tilt camera that can operate in two modes, either narrow field of view or wide field of view using a conic mirror. Fires are detected as hot spots, and the location is detected in order to move the robot to the source of fire. [58] proposes a hybrid system for forest fire detection composed of both thermal and visual cameras, and meteorological and geographical information, while [59] proposes a handheld thermal imaging system for airborne fire analysis.

[60] presents a gunfire detection and localisation system for military applications. Gunfire is detected in mid-wave infrared images and validated by acoustic events. The detected gunfire is mapped to a real-world location. [61] proposes using thermal imaging for mine detection. Depending on circumstances such as the ambient air temperature and soil moisture, mines can be detected using the assumption that the soil directly above the mine heats or

cools at a slightly different rate than the surrounding soil. [62] uses the same idea in their system. They spray cold water over the surrounding soil, and capture the temperature distribution of the cooling soil with a thermal camera. [63] presents the idea of using thermal imaging for detecting snipers. The muzzle flash, the bullet in flight, and the sniper body can be detected.

2.4.3 Humans

In computer vision research, humans are often the subjects observed. Its application areas are very wide, from surveillance through entertainment to medical diagnostics. While the previously mentioned application areas often use simple image processing algorithms, such as thresholding, or even manual inspection of the images, for human systems there has been more emphasis on robust systems with automatic detection and tracking algorithms. Therefore, this part will also contain information about the specific methods.

Just as described for warm-blooded animals, humans try to maintain a constant body temperature, independent of the temperature of the surroundings. This implies that, when capturing a thermal image, the persons stand out from the background in most environments. Taking advantage of that feature could improve the detection step in many vision systems. If a person is observed from a close distance, information can be extracted about the skin temperature distribution. That can be useful for, e.g., face recognition or medical investigations.

Detection and Tracking of Humans

Detection of humans is the first step in many surveillance applications. General purpose systems should be robust and independent of the environment. The thermal cameras are here often a better choice than a normal visual camera. [64] proposes a system for human detection, based on the extraction of the head region and [65] proposes a detection system that uses background subtraction, gradient information, watershed algorithm and A* search in order to robustly extract the silhouettes. Similar approaches are presented in [66, 67], using Contour Saliency Maps and adaptive filters, while [68] presents a detection method based on the Shape Context Descriptor and Adaboost cascade classifier. A common detection problem is that the surroundings during summer are hotter than or equal to the human temperature. [69] tries to overcome this problem by using Mahalanobis distance between pixel values and edge orientation histograms. [70, 71] use automatic thresholding and a sorting and splitting of blobs in order to detect and count people in sports arenas, see Figure 2.10.

Thermal cameras are very useful for the surveillance and detection of intruders, because of their ability to ‘see’ during the night. For trespasser detection, classification is often based on temperature and simple shape cues. Wong et al. propose two trespasser detection systems, one in which they adjust the camera to detect objects in the temperature range of humans, and then classify the

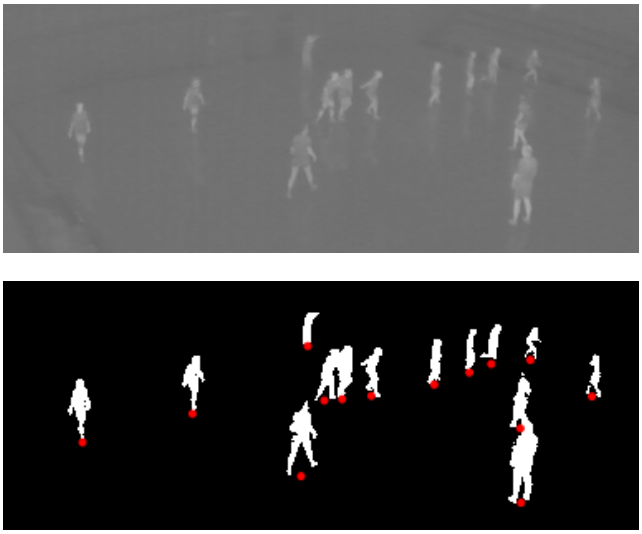


Fig. 2.10: Example of humans playing handball. Top image: Original thermal image. Bottom image: Binarised image with all persons marked with a red dot. [70]

objects based on the shape [72]. The other work aims to identify humans using pattern recognition to detect the human head [73]. [74] uses thresholding, and then a validation of each blob, to determine if it contains one or more persons. If it contains more than one, it will be split into two blobs. [75] proposes a real time detection and tracking system with a classification step based on a cascade of boosted classifiers.

Thermal sensors can be used in systems for the detection of fall accidents or unusual inactivity, which is an important safety tool for the independent living of especially elderly people. [76] proposes a system that uses a low resolution thermal sensor. The system gives an alarm in case of a fall detected, or in the case of inactivity over a long time period. [77] also proposes a fall detection system for private homes by analysing the shape of the detected object. In [78] a fall detection system for bathrooms are proposed, using a thermal sensor mounted above the toilet.

Analysis of more general human activity has also been performed. [79] presents a system that distinguishes between walk and run using spatio-temporal information, while [80] estimates the gait parameters by fitting a 3D kinematic model to the 2D silhouette extracted from the thermal images. In [81] different sports types are classified by the detected location of people over time. [82] proposes a system for analysing the posture of people in crowds, in order to detect people lying down. This could be useful to detect gas attacks or other threats at public places. [83] proposes a system for estimating the human body posture by finding the orientation of the upper body, and locating the major joints of the body.

Rescue robots are used during natural disasters or terrorist attacks, and are often equipped with a thermal camera in order to be able to look for victims in the dark. [84] presents a rescue robot equipped with several sensors, including thermal camera, visual camera and laser range scanner. This robot is able to detect victims in a scene and drive autonomously to the destination. [85] proposes a robot rescue system using thermal and visual camera to identify victims in a scene. For use on Unmanned Aerial Vehicles [86] proposes a human detection system that use a thermal camera to detect warm objects. The shape of the object is analysed in order to reject false detections, before the corresponding region in the colour image is processed with a cascade of boosted classifiers.

Thermal cameras are very popular in the research of pedestrian detection, due to the cameras' independence of lighting changes, which means that it will also work during the night, when most accidents between cars and pedestrians happen. One of the car-based detection systems is proposed in [87], where they present a tracking system for pedestrians. It works well with both still and moving vehicles, but some problems still remain when a pedestrian enters the scene running. [88] proposes a shape-independent pedestrian detection method. Using a thermal sensor with low spatial resolution, [89] builds a robust pedestrian detector by combining three different methods. [90] also proposes a low resolution system for pedestrian detection from vehicles. [91] proposes a pedestrian detection system, that detects people based on their temperature and dimensions and track them using a Kalman filter. In [92] they propose a detection system based on histogram of oriented phase congruency and a SVM classifier for classification of pedestrians. [93] proposes a pedestrian detection system with detection based on symmetric edges, histogram analysis and size of the object. The subsequent work [94] adds a validation step, where the detected objects are matched with a pedestrian model. [95] proposes a system that uses SVM for detection and a combination of Kalman filter prediction and mean shift for tracking.

Wide purpose pedestrian detection includes shape- and appearance-based approaches and local feature-based approaches. [96] uses a shape-based detection and an appearance-based localisation of humans. In [97] the foreground is separated from the background, after that shape cues are used to eliminate non-pedestrian objects, and appearance cues help to locate the exact position of pedestrians. A tracking algorithm is also implemented. [98] uses combinations of local features and classifiers. HOG features and Edgelets are used for detection, and Adaboost and SVM cascade are used as classifiers. [99] and [100] do also use HOG detectors and SVM classifier for pedestrian detection. [101] implements an embedded pedestrian detection system on FPGA. In [102, 103] a car-based stereo-vision system has been tested, detecting warm areas and classify if they are humans, based on distance estimation, size, aspect ratio, and head shape localization. [104, 105] use probabilistic template models for pedestrian classification, while [106] uses a statistical approach for head detection.

For tracking pedestrians, [107, 108] use both spatial and temporal data association, the Wigner distribution, and a motion-based particle filter. [109] uses a multiple-model particle filter, and prior information about the walkways to enhance the performance. [110] does also use a particle filter, combined with two shape- and feature-based measurement models, to track humans in real time from a mobile robot. Other robot-based systems for detection and tracking are proposed in [111, 112]. For the static case, when the robot is still, image differencing and thresholding are applied for human detection. When it moves, the system uses optical flow for filtering the moving foreground objects from moving scene background. [113] proposes a human tracking algorithm for mobile robots that combines a curve matching framework with Kalman filter tracking. [114, 115] propose a local feature (SURF) based method for detection of body parts and tracking of humans. The tracking part uses Kalman-based prediction of object positions to overcome the lack of colour features for distinguishing people. For scenes captured with a moving camera, the Kalman prediction is replaced by a calculation of shift vectors between frames.

Facial Analysis

Face detection is the first step in many applications, including face recognition, head pose analysis, or even some full person detection systems. Since the face is normally not covered by clothes, a thermal camera can capture the direct skin temperature of the face. [116] and [117] propose head detection systems based on a combination of temperature and shape.

Face recognition using thermal cameras eliminates the effects of illumination changes and eases the segmentation step, but it can also introduce some challenges due to the different heat patterns of a subject, caused by different activity levels or emotions such as anxiety. One of the very early approaches is neural networks [118]. [119, 120] compare the use of thermal images in face recognition to visual images using appearance based methods. The thermal images yield better results than visual images here. However, it has not yet been tested how different activity levels of the subjects, and extreme ambient temperature, will affect the recognition rate. In [121] a thermal face recognition algorithm has been developed using the techniques of polar transformation, eigenspace projection, and classification using a multilayer perceptron. [122] tests the use of different parts of the face for facial recognition, and conclude that using the upper part of the face gives a better recognition rate than using the whole face. [123, 124] propose a face recognition system using characteristic and time-invariant physiological information as features.

The recognition of common facial expressions is another task of great interest. Neural networks have also been used as an early approach here [125]. Using a sparse dataset of 120 images, showing four different expressions from one person, the system showed good results. [126] proposes a system to recognise facial expressions by analysing the geometry and local characteristics.

Also facial orientation are of interest in many vision systems. A few papers

have proposed systems to estimate the head pose. [127] calculates the roll angle of frontal face images while [128] proposes a system to estimate the yaw angle of the head. [129] proposes a system for detecting the driver's posture in a car. First the face area is detected, and then the posture is classified as leftward, frontward or rightward.

Measuring the heat distribution in the face can give information about the anxiety level [130], the emotion of car drivers [131], or it can be used for automatic blush detection [132]. For such systems to work automatically, it is important that the system is able to follow the tissue of interest over time. [133] proposes such a tracking system, using a particle filter tracker. For biometric identification, [134] proposes the use of thermal 'faceprints'. These faceprints capture facial physiological features, representing the network of blood vessels under the skin[1]. [139, 140] do also propose the use of thermal face images for biometric identification. They extract the superficial blood vessels from MWIR images with skeletonization. Figure 2.11 shows an example of thermal face and hand images. The veins are visible at the dorsum of the hand.

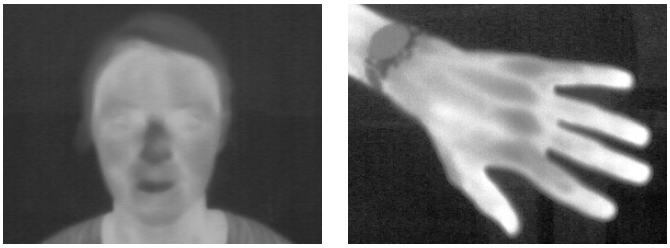


Fig. 2.11: Thermal images of the face or hand can be used for biometric identification.

Medical Analysis

Thermal imaging provides information about physiological processes through examining skin temperature distributions, which can be related to blood perfusion. In the medical area, cameras with high thermal resolution in the human temperature range are used in order to observe fine temperature differences. Thermal imaging complements the standard anatomical investigations based on X-ray and 3D scanning techniques such as CT and MR [141]. [142] and [143] review the medical applications of infrared thermography, including breast cancer detection, diabetes neuropathy, fever screening, dental diagnosis, brain imaging, etc.

Thermal imaging has been shown to reveal tumours in an early state, especially with breast cancer, as described in the survey [144]. Various other medical issues can be studied from thermal images, such as the behaviour of

[1]Thermal images are also used in other biometrics such as hand veins, neck veins and arm veins [135], and palm-dorsa vein patterns [136–138]

the ciliary muscle of the human eye [145], the pulse from the superficial temporal artery [146] or facial vasculature [147], the volumetric flow rate in superficial veins [148], or the periodic fluctuation in skin temperature [149]. Thermal imaging of the human foot has also proven to be useful in the detection of ulceration risks for diabetic patients [150].

In rehabilitation, thermal imaging can be employed to monitor and interpret communications from people with motor impairments, such as mouth opening and closing [151].

2.5 Image Fusion

As discussed in Section 2.1, visual cameras and thermal cameras have both advantages and disadvantages in computer vision applications. Since the limitations of the different technologies are independent, and often do not occur simultaneously, it can be beneficial to combine these different types of images. Occlusion is a well-known problem across all modalities. Separating partly occluded persons or objects of the same temperature can be very difficult in thermal images, as they have the same pixel intensity. Including depth information or colour edges can help disambiguate in this situation.

The most common combination of cameras is thermal and visual. This is due to the low price and well-known characteristics of visual cameras and the ensuing advantages of augmenting colour with temperature.

The main challenges are how to align and fuse the different image modalities. There is not necessarily any relation between brightness level in the different spectra, thus many mutual information alignment methods are not appropriate [152]. Often, corresponding points are manually selected to calculate a planar homography, and then warp one of the images. Automatic alignment techniques that rely on the correlation between edge orientations are presented in [153, 154], and a method that calculates the homography from automatically detected keypoints is presented in [155].

The standard chessboard method for geometric calibration, correction of lens distortion, and alignment of the cameras relies on colour difference, and can not be used for thermal cameras without some changes. [156, 157] report that when heating the board with a flood lamp, the difference in emissivity of the colours will result in an intensity difference in the thermal image. However, a more crisp chessboard pattern can be obtained by constructing a chessboard of two different materials, with large difference in thermal emissivity and/or temperature [158]. This approach is also applied in [159] using a copper plate with milled checker patterns in front of a different base material, and in [160] with a metal wire in front of a plastic board. When these special chessboards are heated by a heat gun, hairdryer or similar, a clear chessboard pattern will be visible in the thermal image, due to the different emissivity of the materials. At the same time, it is also visible in the visual image, due to colour difference. Figure 2.12 shows thermal and RGB pictures from a calibration test. The

chessboard consists of two cardboard sheets, where the white base sheet has been heated right before assembling the board.

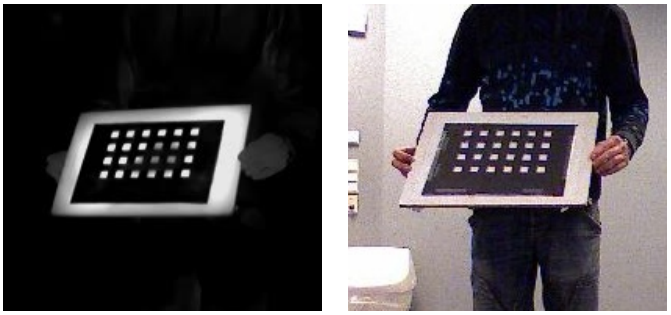


Fig. 2.12: Thermal and visible image of a calibration chessboard consisting of two cardboard sheets, where only the white layer has been heated before assembling.

Fusion can take place at different levels of the processing, often described as pixel level, feature level, or decision level [161]. In pixel level fusion, the images need to be spatially registered, as discussed above, so that the same pixel positions of all images correspond to the same location in the real world. The images are then combined pixel-by-pixel using a fusion algorithm. With feature level fusion, the features are found in all images individually and then fused into a joint feature set. Decision level fusion follows the individual processing until the evaluation of the observed scene is done. At the end-stage, the decisions, or classifications, are combined into one result. The choice of fusion level will depend on the application.

Methods for fusing visible and infrared videos are discussed in [162] and [163], where the two images are aligned and combined using an overlay image with the heat information. In [152] two fusion methods are compared, one named a general fusion model (pixel level) and the other method named a combination module (feature level). The combination module has the best performance tested over six sequences. [164] proposes a combination of curvelet and wavelet transforms, as well as a discrete wavelet packet transform approach for fusing visual and thermal images. In [165] a statistical approach based on expectation maximization is proposed. [166] uses an adaptive weighting method, which enhances unnatural objects in each modality before fusion. An alternative approach is to keep both colour and thermal information as separate channels in a new 'Red-Green-Blue-Thermal' video format [167].

The spatial resolution of thermal cameras is still low, and the price for the best resolution available is high. Using cheap thermal sensors with a low spatial resolution can still improve surveillance when combined with the use of a visual camera. [168] uses a thermal sensor with 16×16 pixels to give an indication of where to search in the visual image. They concentrate on calibrating the two cameras and find the correspondence between the images. By fusing the visual

and thermal videos, super-resolution of the thermal images can be constructed [169]. A prototype using three cameras to combine both the visual, NIR, and LWIR bands is proposed in [170]. Sensors other than visual cameras can also be combined with thermal cameras in fusion systems. [171] fuses the data from a thermal camera and a laser scanner to obtain robust pedestrian detection for tracking. [172] fuses near-infrared sensors and low resolution far-infrared sensors in order to obtain a low-cost night vision system.

Fused image systems find application in a variety of areas, which we now summarize. In surveillance systems it is popular to use fused image modalities, due to the necessity of a robust system that can work both night and day, indoors and outdoors [173, 174]. In both surveillance and night vision systems for cars, the detection of pedestrians can be improved by fusing the image modalities [175–181]. General purpose human detection systems are proposed in [12, 66, 182, 183]. [184] presents a surveillance system that also estimates the attention field of the detected humans and [185] fuses the visual and thermal images for a robust foreground detection in tele-immersive spaces. [186] proposes a combined human detection and recognition system for a robot in domestic environment. Human tracking systems using fused images are proposed in [187–189]. [190] uses visual stereo cameras and a thermal camera, while [191–193] test different combinations of stereo and single cameras of both colour and thermal types.

Face detection and face recognition have been thoroughly investigated and standard methods exist using visual cameras. However, illumination changes still have a negative impact on the performance of these systems [194]. Research has been conducted on whether these problems could be overcome by extending the systems with a thermal sensor. [195] shows a significant improvement by fusing the visible and thermal images in a time-lapse experiment. Work has been done on face recognition using both pixel-level fusion [196–198], feature-level fusion [199], and decision-level fusion [200–202]. [203] proposes fusion on both pixel- and decision-level, while [204] tests two different fusion schemes. They all report an improved performance when fusing the different image modalities.

2.6 Discussion

Although the price of thermal cameras is still significantly higher than the price of comparable visual cameras, the hardware cost is continuously falling and the diversity of cameras is becoming wider. Simple sensors, such as the cheap pyroelectric infrared (PIR) sensor, have for many years been applied as motion detectors for light switch control, burglar alarm, etc. Although no image can be provided by this type of sensor, it can be sufficient for detecting a moving human or large animal. Moving towards thermal cameras, infrared array sensors can read temperature values in a coarse image. These sensors make it possible to analyse the movement, e.g. direction and speed, and can be used for instance in entrance counting systems. The price for these sensors

are less than 50$ for 8×8 pixel arrays with $\pm 2.5°$C temperature accuracy [205]. The price increases with the resolution, framerate, and accuracy, through uncooled cameras to high-end, specialised cooled cameras, with specifications up to 1280×1024 pixels and 130 fps. Some cameras come with even higher framerate, or optical zoom. The price of these high-end cameras can exceed 100,000$.

As seen in this survey, the wide range of cameras opens up for a great diversity of applications of thermal cameras. Each research field has specific needs for, e.g., resolution, field of view, thermal sensitivity, price, or size of the camera. It is therefore expected that the diversity of cameras available will become even larger within the next few years, not only focusing on high-end cameras.

Thermal imaging has found use in two different types of problems: the analysis of known subjects and the detection of unknown subjects. In the first problem, both the subjects and their location in the image are known, and the properties of the subjects can be analysed. The results could be the type of material, condition, or health. The methods used here are often simply the registration of the temperature or even a manual inspection of the images. If computer vision methods are used, often they are just simple algorithms, such as thresholding and blob detection. For the second problem, either the type of objects or their location in the image are unknown. The most important step in this type of problem is normally the detection and classification of objects. The goal here is more often to design an automatic system, e.g., for the detection or tracking of specific objects. More advanced computer vision algorithms can be applied here in order to design a robust and automatic system. In applications where the subject has a different temperature than the surroundings, thermal cameras can significantly ease the detection step compared to visual cameras.

Methods for both analysis of known subjects and detection of unknown subjects are rapidly expanding due to the lower prices and greater availability of thermal cameras. In the case with known subjects, thermal cameras could be viewed as an alternative to a non-contact thermometer. In the last case, the thermal camera is seen more as an alternative to a visual camera, and therefore currently of greater interest from a computer vision point of view. However, the general trend in modern society is the implementation of automation. With this in mind, it is expected that manual and semi-automatic image analysis will gradually be replaced with automatic vision systems, as these become more robust.

The usual disadvantages of changing illumination and the need for active lighting in dark conditions are eliminated with the thermal sensor. Moreover, in the case of surveillance, the use of thermal imaging does not raise as many privacy concerns as the use of visual imaging does. However, new challenges often appear with a new type of sensor. For thermal imaging the lack of texture information can be a disadvantage in some systems, and reflections of the thermal radiation can be a problem in surfaces with high reflectance. For the thermal cameras to stand alone in surveillance purposes, reasonable

priced cameras with higher resolution, effective optical zoom, or wide angle lenses are still desired. In order to overcome some of these challenges it can be advantageous to combine thermal images with other image modalities in many applications. However, there is still a lack of an easy and standardised way to calibrate thermal cameras with other sensors. This must be solved in order to make these types of systems practical. A few pre-calibrated thermal-visual camera set-ups exist today [18, 19], and it is expected to see more of these combined systems in the future.

With more and more sensors becoming available, such as 3D, near-infrared, and thermal, the usual choice of a visual camera is harder to justify. This survey has shown that thermal sensors have advantages in a diversity of applications, and the fusion of different sensors improves the results in some applications. For the future development of vision systems, a careful choice of sensor can open up both new applications as well as alternative features for improving the performance of current applications.

References

[1] MESA Imaging AG. (2012) MESA Imaging SwissRanger. [Online]. Available: http://www.mesa-imaging.ch/index.php

[2] Microsoft. (2012) Kinect. [Online]. Available: http://www.xbox.com/en-US/KINECT

[3] Point Grey Research. (2012) Stereo vision products. [Online]. Available: http://www.ptgrey.com/products/stereo.asp

[4] Sony Electronics Inc. (2012) XC-E150 Near Infrared camera. [Online]. Available: http://pro.sony.com/bbsc/ssr/cat-recmedia/cat-recmediadtwo/product-XCEI50/

[5] J. Byrnes, *Unexploded Ordnance Detection and Mitigation.* Berlin: Springer-Verlag, 2009.

[6] M. Vollmer and K.-P. Möllmann, *Infrared Thermal Imaging—Fundamentals, Research and Applications.* Wiley-VCH, 2010.

[7] Wikipedia. (2006) Atmospheric transmittance. [Online]. Available: http://en.wikipedia.org/wiki/File:Atmosfaerisk_spredning.gif

[8] R. A. Serway and J. W. Jewett, *Physics for Scientists and Engineers with Modern Physics*, 6th ed. Brooks/Cole—Thomson Learning, 2004.

[9] J. D. Hardy, "The radiation of heat from the human body. III. the human skin as a black-body radiator," in *The Journal of Clinical Investigations*, vol. 13(4). American Society for Clinical Investigation, 1934, pp. 615–620.

[10] H. Kaplan, *Practical Applications of Infrared Thermal Sensing and Imaging Equipment*, 3rd ed. SPIE Press, 2007.

[11] FLIR, "Uncooled detectors for thermal imaging cameras," *Tehnical note*, 2011, FLIR Commercial Vision Systems B.V.

[12] J. W. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *Computer Vision and Image Understanding*, vol. 106, no. 2–3, pp. 162–182, 2007.

[13] FLIR, "Cooled versus uncooled cameras for long range surveillance," *Tehnical note*, 2011, FLIR Commercial Vision Systems B.V.

[14] AXIS Communications. (2012) Q1922 datasheet. [Online]. Available: http://www.axis.com/files/datasheet/ds_q1922_q1922-e_46221_en_1207_lo.pdf

[15] FLIR Systems Inc. (2012) FLIR SR-series. [Online]. Available: http://www.flir.com/cs/emea/en/view/?id=41864

[16] Infrared Integrated Systems Ltd. (2012) Thermal imaging cameras by Irisys. [Online]. Available: http://www.irisys.co.uk/thermal-imaging/

[17] Fluke Corporation. (2012) Infrared cameras and thermal cameras by fluke. [Online]. Available: http://www.fluke.com/Fluke/usen/products/category.htm?category=THG-BST&parent=THG

[18] AXIS Communications. (2012) AXIS Network Cameras. [Online]. Available: http://www.axis.com/products/video/camera/index.htm

[19] FLIR Systems Inc. (2012) FLIR product overview. [Online]. Available: http://www.flir.com/cs/emea/en/view/?id=42100

[20] J. Cilulko, P. Janiszewski, M. Bogdaszewski, and E. Szczygielska, "Infrared thermal imaging in studies of wild animals," *European Journal of Wildlife Research*, vol. 59, no. 1, pp. 17–23, 2013.

[21] J. F. Hurnik, S. D. Boer, and A. B. Webster, "Detection of health disorders in dairy cattle utilizing a thermal infrared scanning technique," *Canadian Journal of Animal Science*, vol. 64, no. 4, pp. 1071–1073, 1984.

[22] A. J. Arenas, F. Gómez, R. Salas, P. Carrasco, C. Borge, A. Maldonado, D. J. O'Brien, and F. Martínez-Moreno, "An evaluation of the application of infrared thermal imaging to the tele-diagnosis of sarcoptic mange in the spanish ibex (capra pyrenaica)," *Veterinary Parasitology*, vol. 109, no. 1–2, pp. 111–117, 2002.

[23] M. R. Dunbar and K. A. MacCarthy, "Use of infrared thermography to detect signs of rabies infection in raccoons (procyon lotor)," *Journal of Zoo and Wildlife Medicine*, vol. 37, no. 4, pp. 518–523, 2006.

[24] P. D. Warriss, S. J. Pope, S. N. Brown, L. J. Wilkins, and T. G. Knowles, "Estimating the body temperature of groups of pigs by thermal imaging," *Veterinary Record*, vol. 158, pp. 331–334, 2006.

[25] L. E. Yanmaz, Z. Okumus, and E. Dogan, "Instrumentation of thermography and its applications in horses," *Journal of Animal and Veterinary Advances*, vol. 6, no. 7, pp. 858–862, 2007.

[26] D. Dikic, D. Kolaric, D. Lisicic, V. Benkovic, A. Horvat-Knezevic, Z. Tadic, K. Skolnik-Gadanac, and N. Orsolic, "Use of thermography in studies of thermodynamics in frogs exposed to different ambiental temperatures," in *ELMAR*, 2011.

[27] D. Zhou, M. Dillon, and E. Kwon, "Tracking-based deer vehicle collision detection using thermal imaging," in *IEEE International Conference on Robotics and Biomimetics*, 2009.

[28] FLIR, "BMW incorporates thermal imaging cameras in its cars," *Application story*, 2011, FLIR Commercial Vision Systems B.V.

[29] K. A. Steen, A. Villa-Henriksen, O. R. Therkildsen, H. Karstoft, and O. Green, "Automatic detection of animals using thermal imaging," in *International Conference on Agricultural Engineering*, july 2012.

[30] A. A. Gowen, B. K. Tiwari, P. J. Cullen, K. McDonnell, and C. P. O'Donnell, "Applications of thermal imaging in food quality and safety assessment," *Trends in Food Science and Technology*, vol. 21, no. 4, pp. 190–200, 2010.

[31] R. Vadivambal and D. Jayas, "Applications of thermal imaging in agriculture and food industry—A review," *Food and Bioprocess Technology*, vol. 4, pp. 186–199, 2011.

[32] V. Chelladurai, D. S. Jayas, and N. D. G. White, "Thermal imaging for detecting fungal infection in stored wheat," *Journal of Stored Products Research*, vol. 46, no. 3, pp. 174–179, 2010.

[33] K. Kheiralipour, H. Ahmadi, A. Rajabipour, S. Rafiee, M. Javan-Nikkhah, and D. Jayas, "Development of a new threshold based classification model for analyzing thermal imaging data to detect fungal infection of pistachio kernel," *Agricultural Research*, vol. 2, no. 2, pp. 127–131, 2013.

[34] Public Laboratory. (2012) Thermal photography. [Online]. Available: http://publiclaboratory.org/tool/thermal-photography

[35] A. R. Al-Kassir, J. Fernandez, F. Tinaut, and F. Castro, "Thermographic study of energetic installations," *Applied Thermal Engineering*, vol. 25, no. 2–3, pp. 183–190, 2005.

[36] J. R. Martinez-De Dios and A. Ollero, "Automatic detection of windows thermal heat losses in buildings using UAVs," in *World Automation Congress*, 2006.

[37] L. Hoegner and U. Stilla, "Thermal leakage detection on building facades using infrared textures generated by mobile mapping," in *Joint Urban Remote Sensing Event*, 2009.

[38] D. Iwaszczuk, L. Hoegner, and U. Stilla, "Matching of 3D building models with IR images for texture extraction," in *Joint Urban Remote Sensing Event*, 2011.

[39] B. Sirmacek, L. Hoegner, and U. Stilla, "Detection of windows and doors from thermal images by grouping geometrical features," in *Joint Urban Remote Sensing Event*, 2011.

[40] P. Angaitkar, K. Saxena, N. Gupta, and A. Sinhal, "Enhancement of infrared image for roof leakage detection," in *International Conference on Emerging Trends in Computing, Communication and Nanotechnology*, 2013.

[41] Z. Li, W. Yao, S. Lee, C. Lee, and Z. Yang, "Application of infrared thermography technique in building finish evaluation," *Journal of Nondestructive Evaluation*, vol. 19, pp. 11–19, 2000.

[42] K. James and D. Rice, "Finding termites with thermal imaging," in *InfraMation*, 2002.

[43] E. Grinzato, P. Bison, and S. Marinetti, "Monitoring of ancient buildings by the thermal method," *Journal of Cultural Heritage*, vol. 3, no. 1, pp. 21–29, 2002.

[44] J. L. Lerma, C. Mileto, F. Vegas, and M. Cabrelles, "Visible and thermal IR documentation of a masonry brickwork building," in *CIPA XXI International Symposium*, vol. XXXVI-5/C53. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, october 2007, pp. 456–459.

[45] W. L. Fehlman and M. K. Hinders, *Mobile Robot Navigation with Intelligent Infrared Image*. Berlin: Springer-Verlag, 2010.

[46] J. Meneses, S. Briz, A. J. de Castro, J. Melendez, and F. Lopez, "A new method for imaging of carbon monoxide in combustion environments," *Review of Scientific Instruments*, vol. 68, no. 6, pp. 2568–2573, jun 1997.

[47] E. Ohel, S. R. Rotman, D. G. Blumberg, and L. Sagiv, "Anomaly gas remote sensing and tracking using a field-portable imaging thermal radiometric spectrometer," in *IEEE 24th Convention of Electrical and Electronics Engineers in Israel*, 2006.

[48] A. W. Lewis, S. T. S. Yuen, and A. J. R. Smith, "Detection of gas leakage from landfills using infrared thermography—Applicability and limitations," *Waste Management and Research*, vol. 21, no. 5, pp. 436–447, october 2003.

[49] J. Sandsten, P. Weibring, H. Edner, and S. Svanberg, "Real-time gas-correlation imaging employing thermal background radiation," *Opt. Express*, vol. 6, no. 4, pp. 92–103, Feb 2000.

[50] A. J. Prata and C. Bernardo, "Retrieval of volcanic ash particle size, mass and optical depth from a ground-based thermal infrared camera," *Journal of Volcanology and Geothermal Research*, vol. 186, no. 1–2, pp. 91–107, 2009.

[51] R.-D. Rogler, H. Lobl, and J. Schmidt, "A diagnostic system for live electrical joints in power transmission systems," in *Forty-Second IEEE Holm Conference on Electrical Contacts. Joint with the 18th International Conference on Electrical Contacts*, 1996.

[52] M. S. Jadin, K. H. Ghazali, and S. Taib, "Thermal condition monitoring of electrical installations based on infrared image analysis," in *Saudi International Electronics, Communications and Photonics Conference*, 2013.

[53] L. H. Meyer, S. H. Jayaram, and E. A. Cherney, "A novel technique to evaluate the erosion resistance of silicone rubber composites for high voltage outdoor insulation using infrared laser erosion," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 12, no. 6, pp. 1201–1208, dec. 2005.

[54] Y.-M. H. Ng, M. Yu, Y. Huang, and R. Du, "Diagnosis of sheet metal stamping processes based on 3-D thermal energy distribution," *IEEE Transactions on Automation Science and Engineering*, vol. 4, no. 1, pp. 22–30, jan. 2007.

[55] Z. Hu, Z. Xie, Y. Ci, and W. Wei, "Molten steel level measuring method by thermal image analysis in tundish," in *Recent Advances in Computer Science and Information Engineering*, ser. Lecture Notes in Electrical Engineering. Springer Berlin Heidelberg, 2012, vol. 129, pp. 361–367.

[56] G. Danese, M. Giachero, F. Leporati, N. Nazzicari, and M. Nobis, "An embedded acquisition system for remote monitoring of tire status in F1 race cars through thermal images," in *11th EUROMICRO Conference on Digital System Design Architectures, Methods and Tools*, 2008.

[57] J.-H. Hwang, S. Jun, S.-H. Kim, D. Cha, K. Jeon, and J. Lee, "Novel fire detection device for robotic fire fighting," in *International Conference on Control Automation and Systems*, 2010.

[58] B. C. Arrue, A. Ollero, and J. R. Matinez de Dios, "An intelligent system for false alarm reduction in infrared forest-fire detection," *IEEE Intelligent Systems and their Applications*, vol. 15, no. 3, pp. 64–73, may/jun 2000.

[59] R. Paugam, M. Wooster, and G. Roberts, "Use of handheld thermal imager data for airborne mapping of fire radiative power and energy and flame front rate of spread," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 6, pp. 3385–3399, 2013.

[60] J. Price, C. Maraviglia, W. Seisler, E. Williams, and M. Pauli, "System capabilities, requirements and design of the GDL gunfire detection and location system," in *International Symposium on Information Theory*, 2004.

[61] R. Siegel, "Land mine detection," *IEEE Instrumentation Measurement Magazine*, vol. 5, no. 4, pp. 22–28, dec 2002.

[62] K. Wasaki, N. Shimoi, Y. Takita, and P. N. Kawamoto, "A smart sensing method for mine detection using time difference IR images," in *International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2001.

[63] M. Kastek, R. Dulski, P. Trzaskawka, T. Sosnowski, and H. Madura, "Concept of infrared sensor module for sniper detection system," in *35th International Conference on Infrared Millimeter and Terahertz Waves*, 2010.

[64] T. T. Zin, H. Takahashi, and H. Hama, "Robust person detection using far infrared camera for image fusion," in *Second International Conference on Innovative Computing, Information and Control*, 2007.

[65] J. W. Davis and V. Sharma, "Robust detection of people in thermal imagery," in *Proceedings of the 17th International Conference on Pattern Recognition*, 2004.

[66] J. W. Davis and M. A. Keck, "A two-stage template approach to person detection in thermal imagery," in *Seventh IEEE Workshops on Application of Computer Vision*, 2005.

[67] Z. Li, J. Zhang, Q. Wu, and G. Geers, "Feature enhancement using gradient salience on thermal image," in *International Conference on Digital Image Computing: Techniques and Applications*, 2010.

[68] W. Wang, J. Zhang, and C. Shen, "Improved human detection and classification in thermal images," in *17th IEEE International Conference on Image Processing*, 2010.

[69] A. Jo, G.-J. Jang, Y. Seo, and J.-S. Park, "Performance improvement of human detection using thermal imaging cameras based on mahalanobis distance and edge orientation histogram," in *Information Technology Convergence*, ser. Lecture Notes in Electrical Engineering, 2013, vol. 253, pp. 817–825.

[70] R. Gade, A. Jørgensen, and T. B. Moeslund, "Occupancy analysis of sports arenas using thermal imaging," in *Proceedings of the International Conference on Computer Vision and Applications*, 2012.

[71] ——, "Long-term occupancy analysis using graph-based optimisation in thermal imagery," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[72] W. K. Wong, P. N. Tan, C. K. Loo, and W. S. Lim, "An effective surveillance system using thermal camera," in *International Conference on Signal Acquisition and Processing*, 2009.

[73] W. K. Wong, Z. Y. Chew, C. K. Loo, and W. S. Lim, "An effective trespasser detection system using thermal camera," in *Second International Conference on Computer Research and Development*, 2010.

[74] A. Fernández-Caballero, J. C. Castillo, J. Serrano-Cuerda, and S. Maldonado-Bascón, "Real-time human segmentation in infrared videos," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2577 – 2584, 2011.

[75] Y. Benezeth, B. Emile, H. Laurent, and C. Rosenberger, "A real time human detection system based on far infrared vision," in *Image and Signal Processing*, ser. Lecture Notes in Computer Science. Berlin: Springer-Verlag, 2008, vol. 5099, pp. 76–84.

[76] A. Sixsmith and N. Johnson, "A smart sensor to detect the falls of the elderly," *IEEE Pervasive Computing*, vol. 3, no. 2, pp. 42–47, april-june 2004.

[77] W. K. Wong, H. L. Lim, C. K. Loo, and W. S. Lim, "Home alone faint detection surveillance system using thermal camera," in *Second International Conference on Computer Research and Development*, 2010.

[78] S. Kido, T. Miyasaka, T. Tanaka, T. Shimizu, and T. Saga, "Fall detection in toilet rooms using thermal imaging sensors," in *IEEE/SICE International Symposium on System Integration*, 2009.

[79] J. Han and B. Bhanu, "Human activity recognition in thermal infrared imagery," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2005.

[80] B. Bhanu and J. Han, "Kinematic-based human motion analysis in infrared sequences," in *Sixth IEEE Workshop on Applications of Computer Vision*, 2002.

[81] R. Gade and T. B. Moeslund, "Sports type classification using signature heatmaps," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013.

[82] Q.-C. Pham, L. Gond, J. Begard, N. Allezard, and P. Sayd, "Real-time posture analysis in a crowd using thermal imaging," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[83] S. Iwasawa, K. Ebihara, J. Ohya, and S. Morishima, "Real-time estimation of human body posture from monocular thermal images," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1997.

[84] H. Mano, K. Kon, N. Sato, M. Ito, H. Mizumoto, K. Goto, R. Chatterjee, and F. Matsuno, "Treaded control system for rescue robots in indoor environment," in *IEEE International Conference on Robotics and Biomimetics*, 2009.

[85] M. Z. Aziz and B. Mertsching, "Survivor search with autonomous UGVs using multimodal overt attention," in *IEEE International Workshop on Safety Security and Rescue Robotics*, july 2010.

[86] P. Rudol and P. Doherty, "Human body detection and geolocalization for UAV search and rescue missions using color and thermal imagery," in *IEEE Aerospace Conference*, 2008.

[87] E. Binelli, A. Broggi, A. Fascioli, S. Ghidoni, P. Grisleri, T. Graf, and M. Meinecke, "A modular tracking system for far infrared pedestrian recognition," in *IEEE Intelligent Vehicles Symposium*, 2005.

[88] Y. Fang, K. Yamada, Y. Ninomiya, B. K. P. Horn, and I. Masaki, "A shape-independent method for pedestrian detection with far-infrared images," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 6, pp. 1679–1697, nov. 2004.

[89] M. Mahlisch, M. Oberlander, O. Lohlein, D. Gavrila, and W. Ritter, "A multiple detector approach to low-resolution FIR pedestrian recognition," in *IEEE Intelligent Vehicles Symposium*, 2005.

[90] J.-E. Kallhammer, D. Eriksson, G. Granlund, M. Felsberg, A. Moe, B. Johansson, J. Wiklund, and P.-E. Forssen, "Near zone pedestrian detection using a low-resolution FIR sensor," in *IEEE Intelligent Vehicles Symposium*, 2007.

[91] D. Olmeda, A. de la Escalera, and J. M. Armingol, "Detection and tracking of pedestrians in infrared images," in *3rd International Conference on Signals, Circuits and Systems*, 2009.

[92] D. Olmeda, A. de la Escalera, and J. Armingol, "Contrast invariant features for human detection in far infrared images," in *IEEE Intelligent Vehicles Symposium*, 2012.

[93] M. Bertozzi, A. Broggi, P. Grisleri, T. Graf, and M. Meinecke, "Pedestrian detection in infrared images," in *IEEE Intelligent Vehicles Symposium*, june 2003.

[94] M. Bertozzi, A. Broggi, A. Fascioli, T. Graf, and M.-M. Meinecke, "Pedestrian detection for driver assistance using multiresolution infrared vision," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 6, pp. 1666–1678, nov. 2004.

[95] F. Xu, X. Liu, and K. Fujimura, "Pedestrian detection and tracking with night vision," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 1, pp. 63–71, march 2005.

[96] C. Dai, Y. Zheng, and X. Li, "Layered representation for pedestrian detection and tracking in infrared imagery," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop*, 2005.

[97] ——, "Pedestrian detection and tracking in infrared imagery using shape and appearance," *Computer Vision and Image Understanding*, vol. 106, no. 2-3, pp. 288–299, May 2007.

[98] L. Zhang, B. Wu, and R. Nevatia, "Pedestrian detection in infrared images based on local shape features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[99] F. Suard, A. Rakotomamonjy, A. Bensrhair, and A. Broggi, "Pedestrian detection using infrared images and histograms of oriented gradients," in *IEEE Intelligent Vehicles Symposium*, 2006.

[100] W. Li, D. Zheng, T. Zhao, and M. Yang, "An effective approach to pedestrian detection in thermal imagery," in *Eighth International Conference on Natural Computation*, 2012.

[101] R. Walczyk, A. Armitage, and T. D. Binnie, "An embedded real-time pedestrian detection system using an infrared camera," in *IET Irish Signals and Systems Conference*, 2009.

[102] M. Bertozzi, A. Broggi, A. Lasagni, and M. D. Rose, "Infrared stereo vision-based pedestrian detection," in *IEEE Intelligent Vehicles Symposium*, 2005.

[103] M. Bertozzi, A. Broggi, C. Caraffi, M. D. Rose, M. Felisa, and G. Vezzoni, "Pedestrian detection by means of far-infrared stereo vision," *Computer Vision and Image Understanding*, vol. 106, no. 2–3, pp. 194–204, 2007.

[104] M. Bertozzi, A. Broggi, C. H. Gomez, R. I. Fedriga, G. Vezzoni, and M. Del Rose, "Pedestrian detection in far infrared images based on the use of probabilistic templates," in *IEEE Intelligent Vehicles Symposium*, june 2007.

[105] H. Nanda and L. Davis, "Probabilistic template based pedestrian detection in infrared videos," in *IEEE Intelligent Vehicle Symposium,*, 2002.

[106] U. Meis, M. Oberlander, and W. Ritter, "Reinforcing the reliability of pedestrian detection in far-infrared sensing," in *IEEE Intelligent Vehicles Symposium*, 2004.

[107] C. N. Padole and L. A. Alexandre, "Wigner distribution based motion tracking of human beings using thermal imaging," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2010.

[108] ——, "Motion based particle filter for human tracking with thermal imaging," in *3rd International Conference on Emerging Trends in Engineering and Technology*, 2010.

[109] P. Skoglar, U. Orguner, D. Törnqvist, and F. Gustafsson, "Pedestrian tracking with an infrared sensor using road network information," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, pp. 1–18, 2012.

[110] A. Treptow, G. Cielniak, and T. Duckett, "Real-time people tracking for mobile robots using thermal vision," *Robotics and Autonomous Systems*, vol. 54, no. 9, pp. 729–739, 2006.

[111] A. Fernández-Caballero, J. C. Castillo, J. Martínez-Cantos, and R. Martínez-Tomás, "Optical flow or image subtraction in human detection from infrared camera on mobile robot," *Robotics and Autonomous Systems*, vol. 58, no. 12, pp. 1273–1281, 2010.

[112] J. C. Castillo, J. Serrano-Cuerda, A. Fernández-Caballero, and M. T. López, "Segmenting humans from mobile thermal infrared imagery," in *Proceedings of the 3rd International Work-Conference on The Interplay Between Natural and Artificial Computation: Part II: Bioinspired Applications in Artificial and Natural Computation*. Berlin: Springer-Verlag, 2009.

[113] S. Lee, G. Shah, A. Bhattacharya, and Y. Motai, "Human tracking with an infrared camera using a curve matching framework," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, pp. 1–15, 2012.

[114] K. Jüngling and M. Arens, "Feature based person detection beyond the visible spectrum," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2009.

[115] ——, "Local feature based person detection and tracking beyond the visible spectrum," in *Machine Vision Beyond Visible Spectrum*, ser. Augmented Vision and Reality. Berlin: Springer-Verlag, 2011, vol. 1, pp. 3–32.

[116] S. Krotosky, S. Cheng, and M. Trivedi, "Face detection and head tracking using stereo and thermal infrared cameras for "smart" airbags: A comparative analysis," in *The 7th International IEEE Conference on Intelligent Transportation Systems*, 2004.

[117] J. Mekyska, V. Espinosa-Duro and, and M. Faundez-Zanuy, "Face segmentation: A comparison between visible and thermal images," in *IEEE International Carnahan Conference on Security Technology*, 2010.

[118] Y. Yoshitomi, T. Miyaura, S. Tomita, and S. Kimura, "Face identification using thermal image processing," in *6th IEEE International Workshop on Robot and Human Communication*, 1997.

[119] D. A. Socolinsky, L. B. Wolff, J. D. Neuheisel, and C. K. Eveland, "Illumination invariant face recognition using thermal infrared imagery," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001.

[120] L. Wolff, D. Socolinsky, and C. Eveland, "Face recognition in the thermal infrared," in *Computer Vision Beyond the Visible Spectrum*, ser. Advances in Pattern Recognition. Berlin: Springer-Verlag, 2005, pp. 167–191.

[121] M. K. Bhowmik, D. Bhattacharjee, M. Nasipuri, D. K. Basu, and M. Kundu, "Classification of polar-thermal eigenfaces using multilayer perceptron for human face recognition," in *IEEE Region 10 and the Third international Conference on Industrial and Information Systems*, 2008.

[122] T. Gault, E. Mostafa, A. Farag, and A. Farag, "Less is more: Cropping to improve facial recognition with thermal images," in *International Conference on Multimedia Technology*, july 2011.

[123] P. Buddharaju, I. T. Pavlidis, and P. Tsiamyrtzis, "Pose-invariant physiological face recognition in the thermal infrared spectrum," in *Conference on Computer Vision and Pattern Recognition Workshop*, 2006.

[124] P. Buddharaju, I. T. Pavlidis, P. Tsiamyrtzis, and M. Bazakos, "Physiology-based face recognition in the thermal infrared spectrum," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 613–626, april 2007.

[125] Y. Yoshitomi, N. Miyawaki, S. Tomita, and S. Kimura, "Facial expression recognition using thermal image processing and neural network," in *6th IEEE International Workshop on Robot and Human Communication*, 1997.

[126] G. Jiang and L. Kang, "Character analysis of facial expression thermal image," in *IEEE/ICME International Conference on Complex Medical Engineering*, 2007.

[127] S. Wu, L. Jiang, S. Xie, and A. C. B. Yeo, "A robust method for detecting facial orientation in infrared images," *Pattern Recognition*, vol. 39, no. 2, pp. 303–309, 2006.

[128] X. Yu, W. K. Chua, L. Dong, K. E. Hoe, and L. Li, "Head pose estimation in thermal images for human and robot interaction," in *2nd International Conference on Industrial Mechatronics and Automation*, may 2010.

[129] T. Kato, T. Fujii, and M. Tanimoto, "Detection of driver's posture in the car by using far infrared camera," in *IEEE Intelligent Vehicles Symposium*, 2004.

[130] I. Pavlidis, J. Levine, and P. Baukol, "Thermal image analysis for anxiety detection," in *International Conference on Image Processing*, 2001.

[131] A. Kolli, A. Fasih, F. Al Machot, and K. Kyamakya, "Non-intrusive car driver's emotion recognition using thermal camera," in *Joint 3rd Int'l Workshop on Nonlinear Dynamics and Synchronization and 16th Int'l Symposium on Theoretical Electrical Engineering*, 2011.

[132] K. Harmer, S. Yue, K. Guo, K. Adams, and A. Hunter, "Automatic blush detection in "concealed information" test using visual stimuli," in *International Conference of Soft Computing and Pattern Recognition*, 2010.

[133] Y. Zhou, P. Tsiamyrtzis, P. Lindner, I. Timofeyev, and I. Pavlidis, "Spatiotemporal smoothing as a basis for facial tissue tracking in thermal imaging," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 5, pp. 1280–1289, 2013.

[134] M. Akhloufi and A. Bendada, "Thermal faceprint: A new thermal face signature extraction for infrared face recognition," in *Canadian Conference on Computer and Robot Vision*, may 2008.

[135] N. S. Gnee, "A study of hand vein, neck vein and arm vein extraction for authentication," in *7th International Conference on Information, Communications and Signal Processing*, 2009.

[136] C.-L. Lin and K.-C. Fan, "Biometric verification using thermal images of palm-dorsa vein patterns," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 2, pp. 199–213, feb. 2004.

[137] K.-C. Fan and C.-L. Lin, "The use of thermal images of palm-dorsa vein-patterns for biometric verification," in *17th International Conference on Pattern Recognition*, 2004.

[138] R. Wang, G. Wang, Z. Chen, J. Liu, and Y. Shi, "An improved method of identification based on thermal palm vein image," in *Neural Information Processing*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, vol. 7664, pp. 18–24.

[139] A. Guzman, M. Goryawala, and M. Adjouadi, "Generating thermal facial signatures using thermal infrared images," in *IEEE International Conference on Emerging Signal Processing Applications*, 2012.

[140] A. M. Guzman, M. Goryawala, J. Wang, A. Barreto, J. Andrian, N. Rishe, and M. Adjouadi, "Thermal imaging as a biometrics approach to facial signature authentication," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 1, pp. 214–222, 2013.

[141] B. F. Jones and P. Plassmann, "Digital infrared thermal imaging of human skin," *IEEE Engineering in Medicine and Biology Magazine*, vol. 21, no. 6, pp. 41–48, nov.-dec. 2002.

[142] B. Lahiri, S. Bagavathiappan, T. Jayakumar, and J. Philip, "Medical applications of infrared thermography: A review," *Infrared Physics & Technology*, vol. 55, no. 4, pp. 221–235, 2012.

[143] E. F. J. Ring and K. Ammer, "Infrared thermal imaging in medicine," *Physiological Measurement*, vol. 33, no. 3, p. R33, 2012.

[144] H. Qi and N. A. Diakides, "Thermal infrared imaging in early breast cancer detection-a survey of recent research," in *25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2003.

[145] B. Harangi, T. Csordás, and A. Hajdu, "Detecting the excessive activation of the ciliaris muscle on thermal images," in *IEEE 9th International Symposium on Applied Machine Intelligence and Informatics*, 2011.

[146] S. Y. Chekmenev, A. A. Farag, and E. A. Essock, "Thermal imaging of the superficial temporal artery: An arterial pulse recovery model," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[147] T. R. Gault and A. A. Farag, "A fully automatic method to extract the heart rate from thermal video," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013.

[148] A. Mahmoud, A. EL-Barkouky, H. Farag, J. Graham, and A. Farag, "A non-invasive method for measuring blood flow rate in superficial veins from a single thermal image," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013.

[149] K. Kondo, N. Kakuta, T. Chinzei, Y. Nasu, T. Suzuki, T. Saito, A. Wagatsuma, H. Ishigaki, and K. Mabuchi, "Thermal rhythmography—Topograms of the spectral analysis of fluctuations in skin temperature," in *23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2001.

[150] H. Peregrina-Barreto, L. Morales-Hernandez, J. Rangel-Magdaleno, and P. Vazquez-Rodriguez, "Thermal image processing for quantitative determination of temperature variations in plantar angiosomes," in *IEEE International Instrumentation and Measurement Technology Conference*, 2013.

[151] N. Memarian, T. Chau, and A. N. Venetsanopoulos, "Application of infrared thermal imaging in rehabilitation engineering: Preliminary results," in *IEEE Toronto International Conference Science and Technology for Humanity*, 2009.

[152] C. Ó Conaire, N. O'Connor, E. Cooke, and A. Smeaton, "Comparison of fusion methods for thermo-visual surveillance tracking," in *9th International Conference on Information Fusion*, 2006.

[153] M. Irani and P. Anandan, "Robust multi-sensor image alignment," in *Sixth International Conference on Computer Vision*, 1998.

[154] R. Istenic, D. Heric, S. Ribaric, and D. Zazula, "Thermal and visual image registration in hough parameter space," in *14th International Workshop on Systems, Signals and Image Processing and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services*, 2007.

[155] S. Sonn, G.-A. Bilodeau, and P. Galinier, "Fast and accurate registration of visible and infrared videos," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013.

[156] S. Y. Cheng, S. Park, and M. Trivedi, "Multiperspective thermal ir and video arrays for 3d body tracking and driver activity analysis," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2005.

[157] S. Prakash, P. Y. Lee, T. Caelli, and T. Raupach, "Robust thermal camera calibration and 3d mapping of object surface temperatures," 2006.

[158] S. Vidas, R. Lakemond, S. Denman, C. Fookes, S. Sridharan, and T. Wark, "A mask-based approach for the geometric calibration of thermal-infrared cameras," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 6, pp. 1625–1635, june 2012.

[159] V. Hilsenstein, "Surface reconstruction of water waves using thermographic stereo imaging," in *Proceedings of Image and Vision Computing New Zealand*, 2005.

[160] Yiu-Ming, H. Ng, and R. Du, "Acquisition of 3d surface temperature distribution of a car body," in *IEEE International Conference on Information Acquisition*, 2005.

[161] T. T. Zin, H. Takahashi, T. Toriu, and H. Hama, "Fusion of infrared and visible images for robust person detection," in *Image Fusion.* InTech, 2011, pp. 239–264.

[162] N. D. Rasmussen, B. S. Morse, M. A. Goodrich, and D. Eggett, "Fused visible and infrared video for use in wilderness search and rescue," in *Workshop on Applications of Computer Vision*, 2009.

[163] P. Sissinto and J. Ladeji-Osias, "Fusion of infrared and visible images using empirical mode decomposition and spatial opponent processing," in *IEEE Applied Imagery Pattern Recognition Workshop*, 2011.

[164] P. Shah, S. N. Merchant, and U. B. Desai, "Fusion of surveillance images in infrared and visible band using curvelet, wavelet and wavelet packet transform," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 08, no. 02, pp. 271–292, 2010.

[165] S. Chen and H. Leung, "An EM-CI based approach to fusion of IR and visual images," in *12th International Conference on Information Fusion*, 2009.

[166] E. Lallier and M. Farooq, "A real time pixel-level based image fusion via adaptive weight averaging," in *Third International Conference on Information Fusion*, 2000.

[167] L. St-Laurent, X. Maldague, and D. Prevost, "Combination of colour and thermal sensors for enhanced object detection," in *10th International Conference on Information Fusion*, 2007.

[168] G. D. Jones, M. A. Hodgetts, R. E. Allsop, N. Sumpter, and M. A. Vicencio-Silva, "A novel approach for surveillance using visual and thermal images," in *A DERA/IEE Workshop on Intelligent Sensor Processing*, feb. 2001, pp. 9/1–9/19.

[169] G. Jones, C. Harding, and V. Leung, "Fusion of data from visual and low-resolution thermal cameras for surveillance," *IEE Seminar Digests*, no. 10062, pp. 17–17, 2003.

[170] A. Toet and M. A. Hogervorst, "Towards an optimal color representation for multiband nightvision systems," in *12th International Conference on Information Fusion*, 2009.

[171] B. Fardi, U. Schuenert, and G. Wanielik, "Shape and motion-based pedestrian detection in infrared images: A multi sensor approach," in *IEEE Intelligent Vehicles Symposium*, 2005.

[172] R. Schweiger, S. Franz, O. Lohlein, W. Ritter, J.-E. Kallhammer, J. Franks, and T. Krekels, "Sensor fusion to enable next generation low cost night vision systems," *Optical Sensing and Detection*, vol. 7726, no. 1, pp. 772 610–772 620, 2010.

[173] P. Kumar, A. Mittal, and P. Kumar, "Fusion of thermal infrared and visible spectrum video for robust surveillance," in *Computer Vision, Graphics and Image Processing*, ser. Lecture Notes in Computer Science. Berlin: Springer-Verlag, 2006, vol. 4338, pp. 528–539.

[174] C. Ó Conaire, E. Cooke, N. E. O'Connor, N. Murphy, and A. F. Smeaton, "Fusion of infrared and visible spectrum video for indoor surveillance," *6th International Workshop on Image Analysis for Multimedia Interactive Services*, 2005.

[175] H. Torresan, B. Turgeon, C. Ibarra-castanedo, P. Hébert, and X. Maldague, "Advanced surveillance systems: Combining video and thermal imagery for pedestrian detection," in *In Proc. of SPIE, Thermosense XXVI, volume 5405 of SPIE*, 2004, pp. 506–515.

[176] C. Ó Conaire, E. Cooke, N. O'Connor, N. Murphy, and A. Smeaton, "Background modelling in infrared and visible spectrum video for people tracking," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, june 2005, p. 20.

[177] A. Leykin and R. Hammoud, "Robust multi-pedestrian tracking in thermal-visible surveillance videos," in *Conference on Computer Vision and Pattern Recognition Workshops*, 2006.

[178] ——, "Pedestrian tracking by fusion of thermal-visible surveillance videos," *Machine Vision and Applications*, vol. 21, pp. 587–595, 2010.

[179] A. Leykin, Y. Ran, and R. Hammoud, "Thermal-visible video fusion for moving target tracking and pedestrian classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[180] Y. Ran, A. Leykin, and R. Hammoud, "Thermal-visible video fusion for moving target tracking and pedestrian motion analysis and classification," in *Augmented Vision Perception in Infrared*, ser. Advances in Pattern Recognition. Berlin: Springer-Verlag, 2009, pp. 349–369.

[181] M. San-Biagio, M. Crocco, M. Cristani, S. Martelli, and V. Murino, "Low-level multimodal integration on riemannian manifolds for automatic pedestrian detection," in *15th International Conference on Information Fusion*, 2012.

[182] L. Jiang, F. Tian, L. E. Shen, S. Wu, S. Yao, Z. Lu, and L. Xu, "Perceptual-based fusion of IR and visual images for human detection," in *International Symposium on Intelligent Multimedia, Video and Speech Processing*, 2004.

[183] G. Szwoch and M. Szczodrak, "Detection of moving objects in images combined from video and thermal cameras," in *Multimedia Communications, Services and Security*, ser. Communications in Computer and Information Science, 2013, vol. 368, pp. 262–272.

[184] A. Leykin and R. Hammoud, "Real-time estimation of human attention field in LWIR and color surveillance videos," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2008.

[185] M. J. Johnson and P. Bajcsy, "Integration of thermal and visible imagery for robust foreground detection in tele-immersive spaces," in *11th International Conference on Information Fusion*, 2008.

[186] M. Correa, G. Hermosilla, R. Verschae, and J. Ruiz-del Solar, "Human detection and identification by robots using thermal and visual information in domestic environments," *Journal of Intelligent & Robotic Systems*, vol. 66, pp. 223–243, 2012.

[187] A. Torabi, G. Massé, and G.-A. Bilodeau, "An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications," *Computer Vision and Image Understanding*, vol. 116, no. 2, pp. 210–221, 2012.

[188] J. Zhao and S.-c. S. Cheung, "Human segmentation by fusing visible-light and thermal imaginary," in *IEEE 12th International Conference on Computer Vision Workshops*, 2009.

[189] M. Airouche, L. Bentabet, M. Zelmat, and G. Gao, "Pedestrian tracking using color, thermal and location cue measurements: a DSmT-based framework," *Machine Vision and Applications*, vol. 23, pp. 999–1010, 2012.

[190] S. K. Lee, K. McHenry, R. Kooper, and P. Bajcsy, "Characterizing human subjects in real-time and three-dimensional spaces by integrating thermal-infrared and visible spectrum cameras," in *IEEE International Conference on Multimedia and Expo*, 2009.

[191] S. J. Krotosky and M. M. Trivedi, "On color-, infrared-, and multimodal-stereo approaches to pedestrian detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 4, pp. 619–629, dec 2007.

[192] ——, "Person surveillance using visual and infrared imagery," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1096–1105, aug 2008.

[193] M. Bertozzi, A. Broggi, M. Felisa, G. Vezzoni, and M. Del Rose, "Low-level pedestrian detection by means of visible and far infra-red tetravision," in *IEEE Intelligent Vehicles Symposium*, 2006.

[194] X. Zou, J. Kittler, and K. Messer, "Illumination invariant face recognition: A survey," in *First IEEE International Conference on Biometrics: Theory, Applications, and Systems*, 2007.

[195] D. A. Socolinsky and A. Selinger, "Thermal face recognition over time," in *17th International Conference on Pattern Recognition*, 2004.

[196] S. Moon, S. G. Kong, J.-H. Yoo, and K. Chung, "Face recognition with multiscale data fusion of visible and thermal images," in *IEEE International Conference on Computational Intelligence for Homeland Security and Personal Safety*, 2006.

[197] S. Kong, J. Heo, F. Boughorbel, Y. Zheng, B. Abidi, A. Koschan, M. Yi, and M. Abidi, "Multiscale fusion of visible and thermal IR images for illumination-invariant face recognition," *International Journal of Computer Vision*, vol. 71, pp. 215–233, 2007.

[198] M. Bhowmik, B. De, D. Bhattacharjee, D. Basu, and M. Nasipuri, "Multisensor fusion of visual and thermal images for human face identification using different SVM kernels," in *IEEE Long Island Systems, Applications and Technology Conference*, may 2012.

[199] D. A. Socolinsky, A. Selinger, and J. D. Neuheisel, "Face recognition with visible and thermal infrared imagery," *Computer Vision and Image Understanding*, vol. 91, no. 1–2, pp. 72–114, 2003.

[200] F. M. Pop, M. Gordan, C. Florea, and A. Vlaicu, "Fusion based approach for thermal and visible face recognition under pose and expresivity variation," in *9th Roedunet International Conference*, 2010.

[201] V. E. Neagoe, A. D. Ropot, and A. C. Mugioiu, "Real time face recognition using decision fusion of neural classifiers in the visible and thermal infrared spectrum," in *IEEE Conference on Advanced Video and Signal Based Surveillance*, 2007.

[202] X. Chen, P. J. Flynn, and K. W. Bowyer, "IR and visible light face recognition," *Computer Vision and Image Understanding*, vol. 99, no. 3, pp. 332–358, 2005.

[203] J. Heo, S. G. Kong, B. R. Abidi, and M. A. Abidi, "Fusion of visual and thermal signatures with eyeglass removal for robust face recognition," in *Conference on Computer Vision and Pattern Recognition Workshops*, 2004.

[204] G. Bebis, A. Gyaourova, S. Singh, and I. Pavlidis, "Face recognition by fusing thermal infrared and visible imagery," *Image and Vision Computing*, vol. 24, no. 7, pp. 727–742, 2006.

[205] Panasonic. (2013) Infrared array sonsor: Grid-eye. [Online]. Available: http://pewa.panasonic.com/components/built-in-sensors/infrared-array-sensors/grid-eye/

Chapter 3

Summary

This thesis consists of six parts, with this introductory part being the first. The following four parts each covers a research topic, which will briefly be explained in the introduction of each part. Each chapter organised under the parts consists of a previously published text. The last part will conclude this thesis.

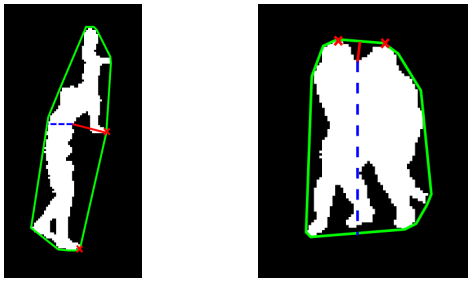
The following sections will give an overview of each chapter, presenting the overall goal, contributions and a summary of results.

3.1 Chapter 4

This chapter consists of the paper "Occupancy Analysis of Sports Arenas Using Thermal Imaging" [1]. This represents the first work towards estimating the occupancy of a sports arena. To goals are defined; estimating the number of people and positions of all people.

The first step in this application is to detect people. Using thermal imaging in an indoor arena simplifies the problem compared to the RGB image modality. Assuming that people are warmer than the background, an automatic thresholding method based on Maximum Entropy [2] is applied. The ultimate goal is to have each person represented by a single white blob (binary large object), but two types of common problems towards this goal must be considered: Groups of people and false detection. For the first issue, groups of people, the aim is to split these connected blobs into smaller blobs consisting of only one person each. In order to do this, the two most common group formations, represented as tall blobs and wide blobs, should be identified. Based on

an initialisation of the camera set-up, the conversion between image and world coordinates is found, as well as the average image height of people, depending on the distance to the camera. White blobs found to be taller than a fixed threshold of 2 meters is assumed to represent two or more people standing behind each other seen from the camera. Analysis of the convex hull and convexity defects of the blob will indicate possible points to split from, illustrated in figure 3.1(a).



(a) (b)

Fig. 3.1: Example of how to find the split location of tall or wide blobs.

Likewise, wide blobs are likely to represent a group of people standing next to each other. These blobs can be identified by the height/width ratio of the blob and the contour length compared to the bounding box perimeter. Blobs satisfying the specified criteria will be split vertically from a point found using a similar shape analysis approach as for the tall blobs. This is illustrated in figure 3.1(b).

False detections can be caused by non-human warm objects, or reflections of infrared waves in glossy surfaces. In the application at hand, most non-human objects are removed by cropping the image to the court area. Furthermore, blobs too small to represent a major part of a human are discarded. Reflections, though, are often observed as heat waves from humans reflected in the floor, pointed towards the camera. In order to identify reflections, the blob is mirrored around a horisontal axis just above its top, and examined if the mirrored blob fits into another existing blob. If this is the case for at least 90 % of the pixels, the blob is classified as reflection and removed.

These algorithms all run as post-processing steps. The remaining blobs are now considered true detections and can be counted in order to estimate the occupancy. For evaluating the spatial occupancy of the court, each person observed in a frame is represented by a 3D Gaussian distribution. The positions of people are calculated by converting the image coordinate of the lower middle point of the bounding box into world coordinates. The occupancy for longer periods can be represented by summing the Gaussian distribution over the given time period.

This work is tested on three weeks of data captured in a sports arena. For

evaluating the precision of the system, 30 minutes of video have been manually annotated with 30 fps and 36 hours have been manually annotated with 1 frame per 25 seconds. For the 30 minute period, an average error in number of people is found to 20.5 %. For the 36 hours of data the algorithm is evaluated for each hour. The results are categorised by the number of people present during that hour. It shows that empty videos are detected with an error close to zero and the error increases with the number of people. This is also expected, as a higher number of people also represents higher complexity of the videos.

3.2 Chapter 5

Chapter 5 consists of the paper "Long-term Occupancy of Analysis using Graph-Based Optimisation in Thermal Imagery" [3]. Aiming for a better occupancy estimation, the main contribution of this paper lies in the graph-based approach optimising between the counting during stable periods and estimated transitions of people.

In this work a new camera set-up is developed in order to capture the entire court of 40×20 meters. Three cameras are mounted in one box and fixed with adjacent fields-of-view. After capturing three video streams of 640×480 pixels simultaneously the images are rectified and stitched to an image of 1920×480 pixels.

Using the algorithms for detecting and dividing blobs described in the previous section, a different sorting step is added. This step is designed to take care of false detections and to combine small blobs that belong to the same person. Based on the idea of probabilistic occupancy maps [4] a candidate rectangle is produced at the position of each detected blob, with the height and width of a standard person at the given position. Each candidate is then evaluated on two criteria; the ratio of white pixels in the rectangle and the ratio of white pixels on the perimeter of the rectangle. The candidate is given a weight based on the two criteria, and for overlapping bounding boxes only the candidate with highest ratio of white pixels is kept.

When evaluation the videos, two different modes can be activated. During so-called stable periods, which are defined by no detections near entrance and exit areas in the image, the number of detected people are simply counted per frame. When people approaches the borders, their position are tracked in order to estimate how many people leave or enter the scene. These two types of information, weighted counts and transitions, are combined in a directed graph, where nodes represent the number of people in stable periods and edges represent the change in number between two periods. An example of the graph is illustrated in fig. 3.2. A dynamic programming approach is applied to calculate the optimal path through the graph.

A full system test is run on 30 minutes manually annotated data captured with 20 fps. The average error per frame is 4.44 %, which is significantly better than both our previous work and comparable methods based on RGB video.

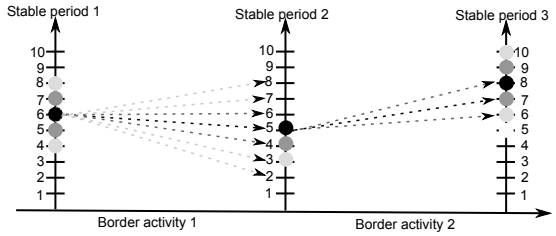


Fig. 3.2: Example of a simple graph. Dark nodes and edges have the highest weight. Edges exist between all nodes in two consecutive periods, but to simplify the illustration they are not drawn.

In addition to the full test, sub-tests are carried out on data captured in five different arenas and on three publicly available thermal video sequences.

3.3 Chapter 6

Continuing to the topic of Activity Recognition, chapter 6 consists of the book chapter "Classification of Sports Types using Thermal Imagery" [5]. In this work the aim is to recognise five different sports types observed in the same indoor sports arena. The main idea of this work is to show the ability to classify sports types based only on position data of all people summed over time. Based on detections of people as described in the previous section, the calibration and conversion of detected positions to world coordinates is of great importance in this work. In order to reduce image distortion, most significantly the barrel effect, a method of calibrating the internal camera parameters is presented. Constructing a calibration board with a grid of 5×4 incandescent light bulbs enables us to adapt traditional "checkerboard" calibration methods. Each of the three cameras is calibrated and images are undistorted individually before manually aligning the cameras to mimic a panoramic view of the arena. Before processing the images they are stitched to one wide image of 1920×480 pixels.

Transformation from image coordinates to world coordinates is usually done using a homography matrix calculated from a set of corresponding coordinates. In the case of three cameras with different viewing angles, a minimum of three homography matrices must be found. This assumes a perfectly rectified image, though. Instead, it is chosen to calibrate the system in a grid of 5×5 meters, producing a higher precision with individual homography matrices for each 5×5 meters square.

After mapping the positions of detected people to world coordinates, they should be represented by a physical area in order to produce a more realistic occupancy map. The representation described in chapter 4 is applied with a 3-dimensional Gaussian distribution of a standard height of 1 and a radius corresponding to 1 metre for 95 % of the volume. The occupancy heatmaps are constructed by adding up the Gaussian distributions over a fixed time interval.

Examples of heatmaps from each sports type are shown in figure 3.3.

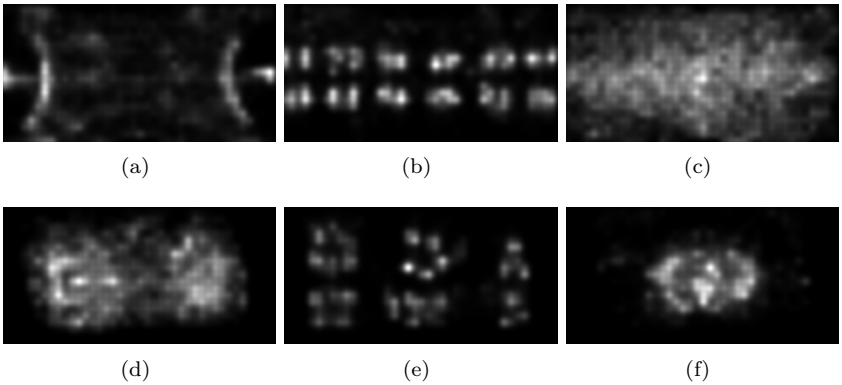


Fig. 3.3: Signature heatmaps of (a) handball, (b) badminton, (c) soccer, (d) basketball, (e) volleyball (three courts), (f) volleyball (one court)

With the hypothesis that an occupancy heatmap image is unique to a specific sports type, within a set of regular sports, classification methods from, e.g., face recognition are considered. Considering each pixel as a dimension in the feature space, PCA is applied to reduce the number of dimensions before Fischer's Linear Discriminant is calculated in the new 20-dimensional space. This method translates the data to the space that best discriminates the different classes of training data [6]. After these transformations, the mean feature values for each class is calculated and used for classifying new images, by simple Euclidean distance in the new space.

From 19 days of video, all hours with a sports activity manually classified as badminton, basketball, soccer, handball or volleyball are picked and separated into a training set and a test set, along with a large amount of videos classified as miscellaneous activities. The result of this test is a correct classification rate of 89.64 %. Additional tests on a uninterrupted full day of video and on a publicly available handball sequence do also very positively approve this method.

3.4 Chapter 7

Chapter 7 consists of the paper "Classification of Sports Types from Tracklets" [7]. Taking a step further towards sports classification independent of scale and rotation of the activities, the aim is here to classify five different sports types from simple motion features. Using a Kalman filter [8] for tracking each player, short trajectories are produced. A trajectory is terminated when no detections is associated for 10 frames, likewise, new trajectories are started when a detection can't be associated with an existing tracker. By this

approach the length of each trajectory will differ. Instead of cropping the trajectories to a fixed length, the track length is used as a feature, representing the complexity of the scenario. The four features chosen for classification is lifespan, total distance, distance span, and mean speed. Lifespan is measured as the number of frames a trajectory exists. The total distance, measured in meters, represents the total distance travelled, as a sum of frame-to-frame distances. The distance span is the maximum distance between any two points of the trajectory, measured in meters. The last feature, mean speed, is found as the mean of the speed between each observation. The four features are chosen from the criteria that they need to be invariant to the size and direction of the court, the position of the players and to the direction of play. They must as well be robust to noisy detections and tracking errors.

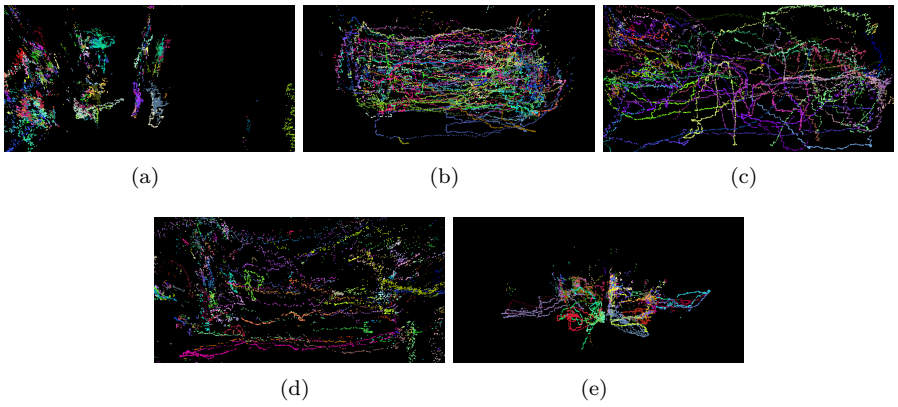


Fig. 3.4: Tracklets from a 2-minute period of (a) badminton, (b) basketball, (c) soccer, (d) handball, and (e) volleyball.

For classification, all videos are split into 2-minutes sequences, for which a single four dimensional feature vector is calculated as mean feature values of all trajectories produced during each sequence. Using a labelled training set, the quadratic discriminant function that best discriminates the data is found. During the classification phase, each new sample is assigned to the class with lowest misclassification cost.

Equivalent to the previous work based on heatmaps, the experiments are performed with five different sports types; Badminton, basketball, handball, soccer and volleyball. 60 minutes of video from each sports type is captured and divided into 2-minutes sequences. Figure 3.4 shows examples of tracklets from a sequence of each sports types. Using a 10-fold cross validation, a correct classification rate of 94.5 % is obtained.

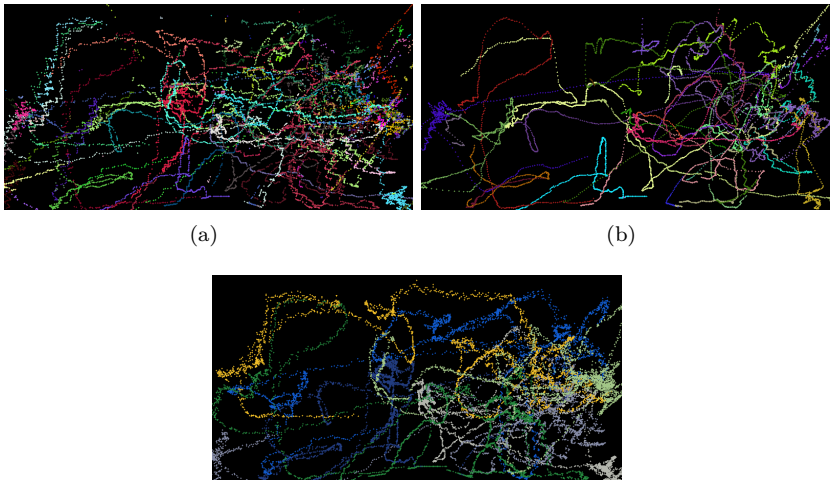
3.5 Chapter 8

Moving into the field of multi-target tracking, chapter 8 consists of the paper "Thermal Tracking of Sports Players" [9]. It presents a tracking framework designed for thermal video, based on the well-known Kalman filter.

The Kalman filter is a recursive algorithm, predicting the next step from previous measurements and associating new measurements to an existing filter, depending on the distance between the predicted step and measurement.

In this work the Kalman filter is applied for multi-target tracking. In order to keep track of several tracks simultaneously, one filter is initialised for each object. In each frame, a set of detections is obtained and must be associated with the existing trackers. Each existing Kalman filter is assigned to the nearest detection, within a given distance threshold. For each detection that is not assigned to a Kalman filter, a new track is started by initialising a new Kalman filter. Kalman filters with no associated detections will be continued by predictions for up to 10 frames, in order handle to short occlusions.

The proposed tracking framework is based on the relatively simple recursive algorithm. A different research direction in multi-target tracking is offline tracking based on batch optimisation. Exploiting the possibility of going back and forth in time, this type of tracking algorithms has produced very good results on standard RGB datasets recently. In the evaluation our work is therefore compared to a publicly available implementation called Continuous Energy Minimization (CEM) [10].



(a) (b)

(c)

Fig. 3.5: Trajectories plot in world coordinates with each identity assigned a random colour. (**a**) Trajectories found by Kalman tracker, (**b**) trajectories found by CEM tracker (60 epochs) and (**c**) manually annotated trajectories.

The test is performed on a challenging thermal video sequence of 2 minutes, capturing an indoor soccer game. The trajectory of each player is manually annotated and the performance of the trackers is evaluated using the CLEAR MOT metrics [11]. Using the same set of detections for both trackers, the Kalman filter is able to produce more tracks, compared to the CEM tracker. The result is an accuracy (MOTA) for the Kalman filter of 70.36 %, while the CEM tracker ends with a negative MOTA value, due to a very high number of false negatives. Figure 3.5 shows the trajectories produced by each method.

The drawback of the Kalman filter is a high number of ID switches. Due to the recursive nature of this algorithm, there is no possibility of connecting broken tracks, and no appearance model, that could be used for re-identification of people, is obtained. The main conclusion of this work is that a Kalman filter is well suited for tracking in the thermal modality. It produces short reliable tracks and it easily performs fast enough for real-time applications.

3.6 Chapter 9

The paper included in chapter 9, "Constrained Multi-Target Tracking for Thermal Imaging", presents the preliminary results for a work in progress. In this chapter the work towards a robust multi-target tracking algorithm for thermal video is continued. The results of the previous chapter showed clearly that the offline tracker tended to produce too few trajectories, while the recursive Kalman filter introduced a high number of ID switches and split tracks. This work will try to improve an offline tracking algorithm for thermal videos by introducing a new term in the energy function, constraining the number of trajectories produced.

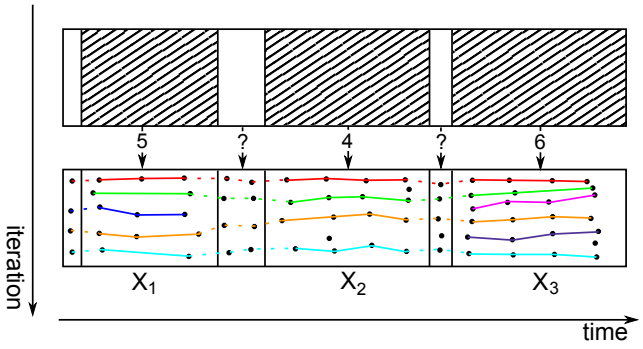


Fig. 3.6: Illustration of the proposed method. During first iteration, stable periods of the video sequence are identified and a number of people present are estimated. This is used as input for the second iteration, in which trajectories are constructed and optimised.

The algorithm runs in two iterations, as illustrated in figure 3.6. During first iteration, periods with a stable number of people present in the scene are

detected, and the number of people is estimated, using the method presented in chapter 5. The results are transferred to the second iteration, in which a global tracker is applied for constructing the trajectories.

For tracking, the starting point is a state-of-the-art publicly available offline tracking algorithm [10], which was also applied in the previous chapter. The purpose of this algorithm is to minimise a global energy function. Included in the energy function is a regularisation term, aiming to minimise the number of targets and maximise the length of each trajectory. In this term the minimisation of targets is discarded and instead a constraint on number is introduced based on the results of the counting algorithm.

The tracking algorithm is tested on four thermal video sequences, three sequences captured from an indoor soccer game and one sequence observing pedestrians in an outdoor courtyard. The results are compared to the original formulation of the tracking algorithm, and shows large improvements from 1-16% in accuracy.

3.7 Chapter 10

Chapter 10 consists of the paper "Improving Global Multi-target Tracking with Local Updates" [12]. This work is conducted during a research stay at the Australian Centre for Visual Technologies (ACVT) in collaboration with local researchers. Therefore, as opposed to the rest of this thesis, this paper focuses on RGB video. However, the focus is still on tracking of sports players and the work presented in this chapter has been tested on two sequences of video from a real match in the Australian Football League (AFL).

This work starts with an offline multi-target tracking algorithm, and aim to solve one of the problems often seen when occlusions appear; interrupted trajectories.

Assuming that people can only enter and leave the scene at the border of the image, every trajectory terminated far from the image border are localised and considered for improvement. A spatio-temporal window around the detection is chosen in order to limit the complexity of the local data association problem, but still allow for different solutions to be tested. Within the spatio-temporal window, a single target visual tracker is initialised for each trajectory involved in the error. Belonging to the class of tracking-by-detection algorithms, the single target tracker usually picks the detection with highest classification score and associates it to the tracker. However, in crowded scenes with ambiguous situations and occlusions, several trackers might detect the same object as the best candidate. A specific data association scheme is proposed to solve this problem. For each tracker, several candidates is considered and added to a bipartite graph, in addition to an occlusion node for each tracker. Using the classification scores and a spatial proximity measure for weighting the edges, the optimal solution of the graph is found using the Hungarian algorithm [13].

The local solution is inserted into the original global solutions and evaluated

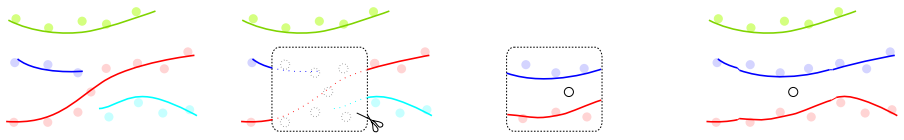


Fig. 3.7: Overview of the optimisation algorithm. Given a possibly erroneous solution (a), we locate each error (b) and perform a local optimisation within its neighbourhood (c). The newly obtained solution is inserted back into the original one if and only if it increases the overall likelihood considering all remaining frames and targets (d).

with the same discrete-continuous objective function as the original solution. If improving the overall energy, the new local solution replaces the original solution. The algorithm is illustrated in figure 3.7.

The proposed method is evaluated on eight different sequences. Six videos are publicly available datasets and enables us to compare with state-of-art multi-target tracking algorithms. The remaining two sequences are captured during a game of Australian Rules Football. These sequences are particularly challenging due to frequent crowding and physical contact between players.

In short, the evaluation shows how the local optimisations improve the results. Our proposed algorithm outperforms comparable recent multi-target tracking algorithms.

3.8 Chapter 11

Moving into part V, the focus will change from sports applications to other applications of thermal imaging, specifically related to the Smart City. Thermal imaging have clear advantages compared to RGB cameras when installed in public spaces. Firstly, privacy is a big issue when capturing video in public, uncontrolled environments. As the thermal sensor doesn't capture sensitive data, people can not be identified from thermal images. Furthermore, capturing outdoor video during day and night might be challenging, or even impossible, with regular RGB cameras, due to changing lighting conditions and darkness during night. In this part of the thesis it is demonstrated how the previously presented methods for detecting and tracking people can easily be adapted and applied in the context of Smart Cities.

Chapter 11 consists of the paper "Thermal Imaging Systems for Real-Time Applications in Smart Cities" [14]. The purpose of this paper is to demonstrate five successful practical applications in which thermal imaging is applied. In order to be applicable in real-time applications, the applied computer vision methods must be fast, but also robust to environmental changes.

The first work focuses on people counting in a pedestrian street. With the aim of robust counting 24 hours a day, the counting method is based on double differencing of the image, and then uses a conversion between pixel activity

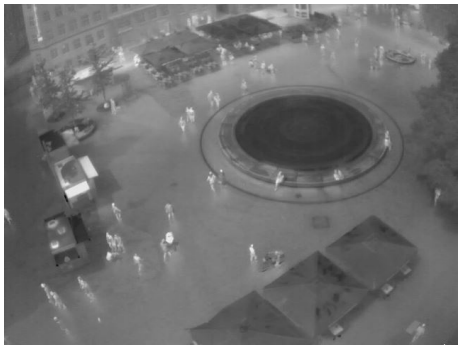


Fig. 3.8: Thermal image overlooking an urban square.

and number of people found from training data. One full week of video has been captured. For evaluation of the precision of the system 13 sequences of 20 minutes have been chosen to cover different days, time of day and level of activity. Using a leave-one-out approach, the mean accuracy of the system is 90.75 %.

The second application focuses on people tracking in a public urban space. The purpose here is to intelligently control the lighting of the space based on the positions and movement of people. The main part of the urban space is covered with three thermal cameras. 16 RGB LED lamps are installed and connected to a computer, which controls the lamps based on the input from computer vision analysis. With experiments running for one week, different lighting scenarios have been tested. Detection of people is based on background subtraction and tracking implemented with a Kalman filter. The system is able to run real-time with 15 fps on a regular laptop computer. This work is described in more detail in the paper presented in chapter 12.

The third application presented in chapter 11 is a tool for evaluating the traffic safety in large urban intersections. Aimed at detecting potential conflicts between cyclists and cars, a thermal camera is installed overlooking the intersection. Using optical flow to detect and estimate the direction of motion in predefined zones of the intersection, the system can flag potential situations of conflicts for further manual investigations. Four different intersections have been used for testing the system. Evaluations show that the false positive rate is high, around 33 %, while only very few false negatives are observed. For the purpose of flagging potential conflict situations for later manual verification, it is desired to allow more false positive and avoid false negatives.

The fourth application focuses on people counting and activity recognition in sports arenas. This section sums up the work presented in chapters 4, 5, and 6 of this thesis.

The last application presented here again applies tracking of people in public spaces. For urban planning and management, knowledge about the human behaviour and movement is important. For large scale studies, the method

must be automatic and fast. It is shown here that computer vision with thermal cameras can be applied for fast and robust detection and tracking of people. Using a georeferenced coordinate system, it is possible to analyse the movement of people in Geographical Information Systems. The paths can be visualised and analysed individually. This work is described in more detail in the paper presented in chapter 13.

3.9 Chapter 12

Chapter 12 consists of the paper "Controlling Urban Lighting by Human Motion Patterns - Results from a Full Scale Experiment" [15]. In this paper we detect and track human movements in a urban square, and use the data as input to the interactive lighting of the square. In this work a real-time solution is presented for detecting and tracking groups of people and converting their positions to a common world coordinate system. People are detected using background subtraction, while tracking is based on Kalman filtering, as previously described in chapter 8.

Fig. 3.9: Two frames showing the "White Aura" scenario, where the light is centred around a person walking across the space.

Four different light scenarios are tested, where two are directly dependent on the detected positions of people. The "White Aura" scenario illuminates a large circle around people. Pictures from this scenario is shown in figure 3.9. The "Treasure Hunt" scenario encourages people to play with the lighting and walk towards a differently coloured lamp, which will then trigger a wave of light through the square.

Tested on cold winter days, the interactive lighting didn't change the behaviour of people. However, observed from the distance, the space seemed more alive and interesting. Tested over one week, the algorithms were proved robust and performed in real-time.

3.10 Chapter 13

The last paper included in this thesis is "Taking the Temperature of Pedestrian Movement in Public Spaces" [16]. Following the same approach of detecting and tracking people as used in chapter 12, the purpose here is different. The trajectories are used as a tool for analysing the use of public space and it is shown that the traditional manual observation methods can be substituted by automatic analysis using computer vision.

The location for the experiments is a busy plaza in Copenhagen, with 50-100 pedestrians observed per minute. In this case an empty background image can't be obtained, and the background is instead modelled as the median image over a 30 seconds initializing period. The positions of detected people are projected to real world GPS coordinates in order to relate the positions to the surroundings and to other geospatial data layers in Geographical Information Systems (GIS).

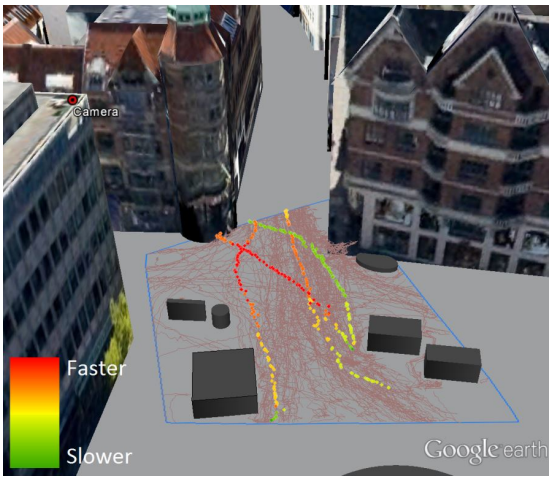


Fig. 3.10: The site and the FOV illustrated in a 3D view with four selected tracks from the same 30 second period coloured according to their speed. All four tracks start in the bottom of the image in the open square and move up towards the street. The tracks indicate both increasing and decreasing speeds. All tracks from a 5-minute period are displayed as background. The position of the camera is shown on the corner of the building to the left. The obstacles are indicated in grey.

During post-processing of the tracking data, statistics of the human movements, such as distance, duration, speed, and start and end points are extracted. In this process a filtering of data is also performed based on the duration of each track. Tracks shorter than 3 seconds are classified as noise and discarded. Most often these short tracks are produced by non-human objects moving in the wind or caused by occlusions between people.

5 minutes of data from one view and 1 minute from a second view have been manually annotated with trajectories of all people and compared to the

automatic results. The results show that the general movement patterns are captured very precisely and the positioning is good. The main problems observed for the automatic tracking system are split tracks and identity switches in cases of occlusions. This is also expected, as the Kalman filter is a recursive algorithm with no re-identification of people, or re-connection of tracks, possible.

3.11 Contributions

This section will sum up the contributions made in this thesis.

- **Survey on thermal cameras and applications.** As an introduction of thermal cameras to the computer vision community, chapter 2 explains the nature of thermal radiation and the technology of thermal cameras. Moreover, a thorough survey of applications of thermal cameras is presented.
- **Detection of people in thermal images**. In chapter 4 and 5 algorithms for detecting individual people in thermal images are presented. This includes methods for splitting and sorting detected objects.
- **Robust counting of people**. The counting method presented in chapter 5 includes temporal information on the transition in occupancy to account for noisy detections.
- **Sports type classification**. Chapter 6 and 7 introduces two different methods for classifying sports types performed in the same arena. One method based only on frame-based position of people, while the other method is based on features extracted from short trajectories.
- **Tracking applied to thermal video**. Few works have been designed for and tested on thermal video. In chapter 8 a real-time multi-target tracking algorithm is presented and chapter 9 introduces a method to improve tracking performance by constraining the number of trajectories produced by an offline tracking algorithm.
- **Demonstration of practical applications of computer vision with thermal imaging**. The methods developed in this thesis have been applied and tested in several real world cases. Part V presents five different applications.

References

[1] R. Gade, A. Jørgensen, and T. B. Moeslund, "Occupancy analysis of sports arenas using thermal imaging," in *Proceedings of the International Conference on Computer Vision and Applications*, vol. 2, feb. 2012, pp. 277–283.

[2] J. Kapur, P. Sahoo, and A. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Computer Vision, Graphics, and Image Processing*, vol. 29, no. 3, pp. 273 – 285, 1985.

[3] R. Gade, A. Jørgensen, and T. Moeslund, "Long-term occupancy analysis using graph-based optimisation in thermal imagery," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013, pp. 3698–3705.

[4] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *PAMI*, vol. 30, no. 2, pp. 267 –282, feb. 2008.

[5] R. Gade and T. B. Moeslund, "Classification of sports types using thermal imagery," in *Computer Vision in Sports*, T. B. Moeslund, G. Thomas, and A. Hilton, Eds. Springer, January 2015.

[6] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, 2001.

[7] R. Gade and T. Moeslund, "Classification of sports types from tracklets," in *KDD workshop on Large-Scale Sports Analytics*, August 2014.

[8] G. Welch and G. Bishop, "An introduction to the kalman filter," Chapel Hill, NC, USA, Tech. Rep., 1995.

[9] R. Gade and T. B. Moeslund, "Thermal tracking of sports players," *Sensors*, vol. 14, no. 8, pp. 13 679–13 691, 2014.

[10] A. Andriyenko and K. Schindler, "Multi-target tracking by continuous energy minimization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011, pp. 1265–1272.

[11] K. Bernardin and R. Stiefelhagen, "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, p. 246309, 2008.

[12] A. Milan, R. Gade, A. Dick, T. B. Moeslund, and I. Reid, "Improving global multi-target tracking with local updates," in *ECCV workshop on Visual Surveillance and Re-Identification*, September 2014.

[13] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, pp. 83–97, 1955.

[14] R. Gade, T. B. Moeslund, S. Z. Nielsen, H. Skov-Petersen, H. J. Andersen, K. Basselbjerg, H. T. Dam, O. B. Jensen, A. Jørgensen, H. Lahrmann, T. K. O. Madsen, E. S. Bala, and B. O. Povey, "Thermal imaging systems for real-time applications in smart cities," *International Journal of Computer Applications in Technology*, accepted for publication.

[15] E. S. Poulsen, H. J. Andersen, O. B. Jensen, R. Gade, T. Thyrrestrup, and T. B. Moeslund, "Controlling urban lighting by human motion patterns - results from a full scale experiment," in *ACM International Conference on Multimedia (MM)*, 2012.

[16] S. Z. Nielsen, R. Gade, T. B. Moeslund, and H. Skov-Petersen, "Taking the temperature of pedestrian movement in public spaces," *Transportation Research Procedia*, vol. 2, pp. 660 – 668, 2014, the Conference on Pedestrian and Evacuation Dynamics 2014 (PED 2014).

Part II

Occupancy analysis

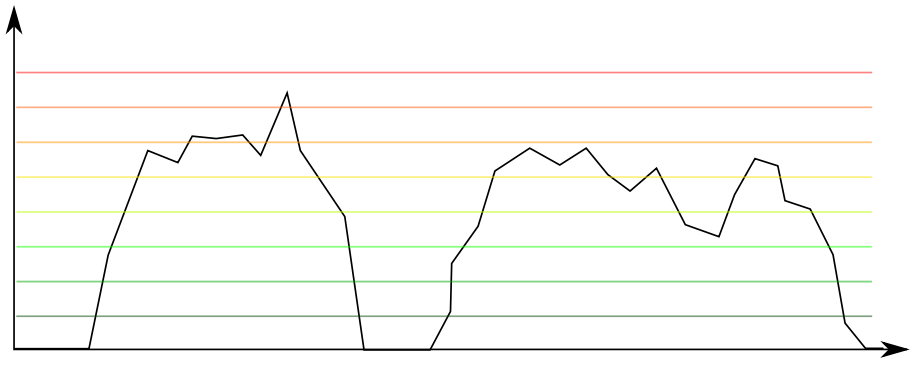
Large facilities, such as sports arenas and cultural centres, are expensive in both building expenses and maintenance. Knowledge about the use and degree of occupancy of these facilities are therefore of great importance in order to optimise the use and include these experiences when planning on new facilities.

Traditionally, occupancy analysis of public facilities has been conducted by manual inspections, either by the administrative staff in the buildings or by external observers. Hiring external observers to watch the activities in each facility over long periods is very expensive. The alternative solution of random checks during a day might be imprecise and misleading for the true occupancy pattern. The work presented in this part suggests two automatic systems based on thermal video feed. A camera can capture data day and night without significant extra expenses. Using the computer vision methods developed in this section the occupancy of a sports facility is automatically estimated.

This part is the first of four main parts of this thesis and includes two published conference papers:

Rikke Gade, Anders Jørgensen and Thomas B. Moeslund, "Occupancy Analysis of Sports Arenas Using Thermal Imaging," *Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP)*, pp. 277–283, February 2012.

Rikke Gade, Anders Jørgensen and Thomas B. Moeslund, "Long-term Occupancy Analysis using Graph-based Optimisation in Thermal Imagery," *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3698–3705, June 2013.



Chapter 4

Occupancy Analysis of Sports Arenas using Thermal Imaging

Rikke Gade, Anders Jørgensen and Thomas B. Moeslund

The paper has been published in the
Proceedings of the International Conference on Computer Vision and Applications (VISAPP), pp. 277–283, February 2012.

© 2012 SciTePress
The layout has been revised.

Abstract

This paper presents a system for automatic analysis of the occupancy of sports arenas. By using a thermal camera for image capturing the number of persons and their location on the court are found without violating any privacy issues. The images are binarised with an automatic threshold method. Reflections due to shiny surfaces are eliminated by analysing symmetric patterns. Occlusions are dealt with through a concavity analysis of the binary regions. The system is tested in five different sports arenas, for more than three full weeks altogether. These tests showed that after a short initialisation routine the system operates independent of the different environments. The system can very precisely distinguish between zero, some or many persons on the court and give a good indication of which parts of the court that has been used.

4.1 Introduction

In the modern world jobs are becoming ever more sedentary and less physically demanding. This leads to higher demands for activities in people's spare time, which puts a still growing pressure on the sports arenas. From 1964 to 2007 the number of athletes has quadrupled with a steady increase [1]. Surveys also show that people are dropping the classic club sports in favour of more flexible sports [2]. This calls for a better and more optimal use of the existing sports arenas to keep up with this growing trend.

In order to improve the utilisation of a sports arena, its existing use must be examined. This includes examining the number of users using the arena at the same time and the occupancy of the court. Administrators are especially interested in whether the arena is empty, used by a few people or full and the time for when the occupancy changes. The position of the users is also important as they might only use half a court, which means the other half could be rented out to another group. Manual registration of this would be cumbersome and expensive and an automatic approach is therefore needed. For such a system to work in general it should be independent of the size of the court, lighting conditions and without any interaction with the users. This can be obtained with a camera.

Detecting people with a camera raises some privacy issues though. Not all people like surveillance and the fear of being observed could keep some people out of the arenas. This work therefore proposes an automatic method to analyse the occupancy of a sports arena using thermal imaging. One of the advantages of thermal cameras is that the persons recorded cannot be identified, which is an important factor if the system is to be accepted by the users of the sports arena. On top of that, thermal cameras are invariant to lighting, changing backgrounds and colours, which make them more desirable for a general application.

4.2 Related Work

Automatic detection and tracking of sports players is a research area important for all sports analysis. Most systems are using visual cameras. In [3] a tracking system is proposed specifically for indoor football players, while [4] proposes a tracking system for outdoor football using multiple cameras. The tracking system proposed in [5] focuses on more general sports video and it is tested on both football, basketball and hockey.

The large research area regarding automatic identification of human subjects and their behaviour include both visual and thermal cameras. There exist a number of surveys and books on the subject, including [6], [7], [8] and [9].

Thermal cameras measure the amount of thermal radiation, which lies in the long-wavelength infrared spectrum (8-15 μm). All objects with a temperature higher than the absolute zero emit thermal radiation. The intensity and dominating wavelength depends on the temperature.

Thermal cameras have a clear advantage over visual cameras in night conditions, therefore the main focus for systems using thermal cameras have been on security applications and trespasser detection. A few papers with the purpose of detecting trespassers include [10] and [11].

Other work using thermal cameras include systems for pedestrian detection and tracking. In [12] a pedestrian detection method is presented based on the Shape Context Descriptor with the Adaboost cascade classifier framework. [13] proposes the pedestrian detection as part of a driver assistant system while [14] proposes a people detection system for different environments based on contour analysis.

Most vision systems, including the systems mentions above, are only tested on very short video sequences, proving the concept in one or few conditions. In this work the most important issue is stability over a long time period and under different conditions. Therefore the system will be tested over three weeks and in five different arenas. The main results will be average values showing the tendency of occupancy for hours or days.

4.3 Methods

4.3.1 System Overview

The desired system should take a thermal grey scale image as input and find every person in the image. In order to analyse the nature of the problems related to this work, five different sports arenas were selected and used to develop and test the system. During the initial investigations some typical difficulties to obtain the result were registered. Some of these difficulties were occlusions and reflections from both persons and other warm objects, e.g. lamps, on the floor. These typical difficulties must all be addressed in order to make a general system.

As the intention is to monitor the long term use of a sports arena, the system should always be operating. Therefore data should be processed in soft real-time to avoid data pile ups. The output of the system should be, for a given time, the number of users on the court and their position.

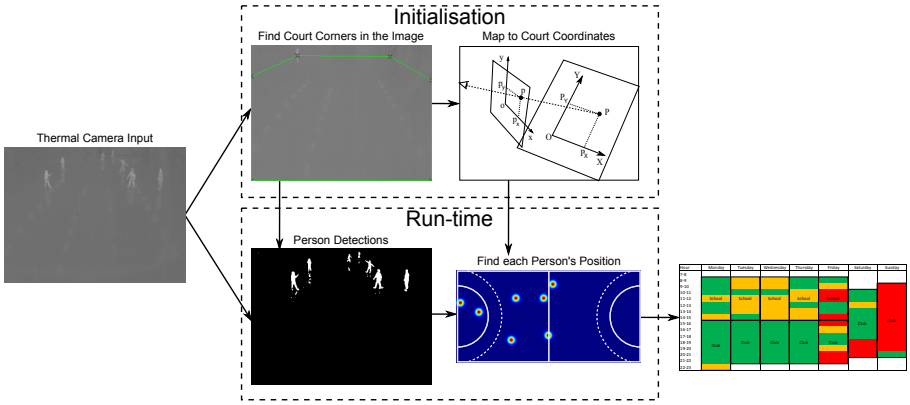


Fig. 4.1: Diagram showing the system steps.

The system should be independent of the camera's viewing angle in relation to the court, as long as the camera can observe the court and is placed in a sufficient height to avoid users covering each other too much. The users' size, level of activity and posture should not have an effect on the measurement either. Figure 4.1 shows a diagram of the system structure.

The overall idea is to develop a system that can detect persons in a thermal image and, with the inputs from the initialisation, find the persons' positions at the court. As mentioned in the introduction, the data should be categorised after the occupancy level into zero, some and many people and presented in a timetable.

4.3.2 Initialising the System

The initialisation routine must be conducted for each new mounting of the camera. This routine handles the adjustments necessary to fit the system to the layout of the actual sports arena. First the court must be found in the image to avoid that cold or hot objects outside the court influences the system. As it is only wanted to measure the players at the court, spectators should also be removed. Defining the court in the images gives the opportunity to remove all objects standing outside the court. During the initialisation the corners of the visible part of the court should be marked in the image, and lines connecting them define the outline of the court. To find the persons' positions at the court a mapping from the image to the court must be found. Using at least four corresponding points in image and world coordinates, a homography matrix $\mathbf{H}$ can be calculated [15].

After initialising this matrix the mapping between image coordinates and world coordinates is calculated as $P_w = Hp_i$, where P_w are the weighted world coordinates $[P_X\ P_Y\ W]^T$ and p_i are the image coordinates $[p_x\ p_y\ 1]^T$. The real world coordinates are found by dividing P_w with the weight W.

At least four corresponding points must be used in order to calculate $\mathbf{H}$, but tests of the homography show that using more points increase the precision. This is due to nonlinearity in the mapping, as the lens has some barrel effect. Therefore it is desirable to use as many points in the initialisation as possible. In this work a two-dimensional grid with steps of 5 metres is used to mark the points at the court. A hot or cold object is necessary to detect the grid points in the image.

4.3.3 Run-Time

This continuous loop receives an image from the thermal camera and after a number of functions it delivers a set of regions each containing one person. First the thermal camera captures a grey scale image. The warm objects (persons) are bright while the surroundings are dark grey. After capturing a frame the first step is to extract the warm objects. For this an automatic threshold method based on Maximum Entropy is used [16]. This method maximises the sum of the entropy above and below the threshold value s, by iterating through every possible value. The threshold function is only run for the pixels inside the court area, to avoid disturbance from spectators. The result is a binary image where ideally the persons are white and anything else is black. If the maximum entropy is below a specified threshold TH there are no persons on the court and the frame can be discarded.

The white regions are now found using the contour finding algorithm described in [17]. If there are no valid regions i.e. regions larger than a specified minimum area the frame is discarded.

Split Tall Regions

People standing behind each other, seen from the camera's point of view, can often be found as one tall region as shown in figure 4.2(a). In order to split such regions into the right number of people, it must be investigated when a region is too high to contain only one person. Using the camera's height c, the vertical resolution r_v and vertical field-of-view f_v of the camera, the height in pixels can be found as a function of the person's height p and distance to the camera x:

$$y_p = \frac{r_v \cdot \left(tan^{-1}\left(\frac{x}{c-p}\right) - tan^{-1}\left(\frac{x}{c}\right)\right)}{f_v}$$

Statistics show that only 0.26% of Danish conscripts were taller than 2 metres [18], therefore 2 metres is chosen as the height limit. So for each region found in the image the distance to the camera is calculated using the homography and

if the pixel height corresponds to more than 2 metres, the algorithm should try to split the region horizontally. This is done by finding the convex hull and the convexity defects of the contour, as shown in figure 4.2(b). The point selected to split from is the defect point with the largest depth and a maximum absolute gradient of 1.5. The gradient is calculated for the line from the defect point perpendicular on the line between the convexity defect start and end points (green points and yellow line in figure 4.2(c)). The defect point between the legs of the person has the largest depth, but is discarded because the gradient is too high. Also the defect point should not be in the top or bottom fourth of the region, to avoid e.g. feet or head to be split from the body. As shown in figure 4.2(d) the region is split horizontally from the selected point.

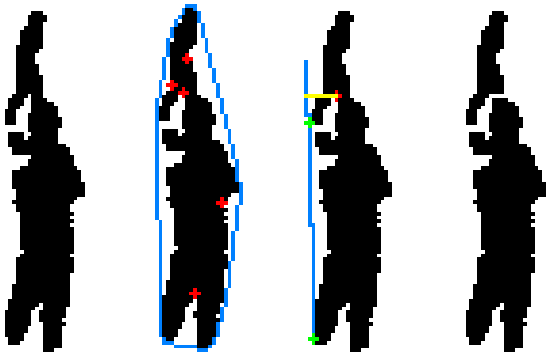


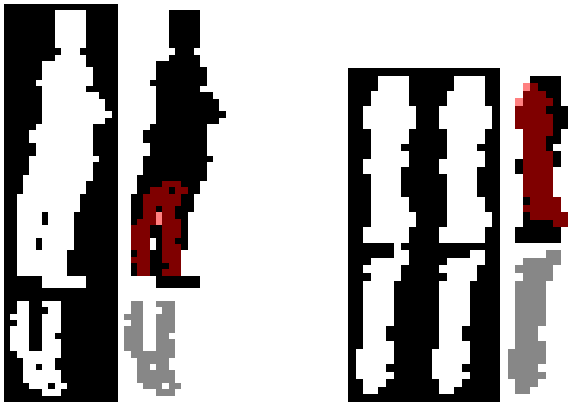
Fig. 4.2: Example on division of tall regions. Note that black and white colours are reversed for better visibility. The blue line is the convex hull, red marks indicate convexity defects and the yellow line the orthogonal depth of the defect.

The algorithm starts with the defect point with largest depth and continues until a point with an acceptable gradient and location is found. If no accepted points are found, the region will not be split. If the region has been split, the algorithm will start over and examine the height of the resulting regions.

Remove Reflections

Just as visible light, infrared waves are also reflected in glossy surfaces, but as the infrared reflections are created by the persons themselves they are always pointing towards the camera. Therefore the mirror axis will be roughly horizontal, and reflections could be removed by trying to mirror them in a region above. An example can be seen in figure 4.3(a)(Left).

In order to remove reflections the system searches for regions that are below a larger or equally sized region. If such a region is found it is mirrored up in the upper region to see if it fits in the person region. If it does not, the reflection is translated one pixel horizontally and checked again. This continues up to three



(a) (b)

Fig. 4.3: Persons having their reflection removed. The red areas mark the reflections after they have been mirrored and translated. In (b) the reflection is first split from the person by the algorithm splitting tall regions.

pixels in all directions. If more than 90 % of the reflection is within the person region it is marked as a reflection and removed. Figure 4.3(a)(Right) shows a situation where 77 out of 79 pixels are within the person region resulting in a coverage of $\approx 97\%$.

In some cases the reflection is connected to the person who created it. See figure 4.3(b)(Left). In these situations the region should first be split by the function splitting tall regions. Figure 4.3(b) shows a situation where a region is first split and secondly the reflection can be removed. Here 72 of 74 ($\approx 97\%$) reflection pixels are within the person.

Split Wide Regions

People standing close to each other will often form one large region. In order to count the people correct such regions must be divided into regions containing only one person. For groups of people standing side by side, seen from the camera's point of view, it will often be possible to separate them based on their head position. Since their heads are narrower than the body they can often be separated by cutting vertically from the minimum points of the upper edge.

As it is not desired to split regions containing only one person, two criteria for the regions must be satisfied before looking for a minimum point to split from. Measuring the features of several regions gives the criteria that to contain more than one person the height of the bounding box must be less than five times the width and the contour of the region must be longer than the bounding box perimeter. If these criteria are satisfied and a minimum point can be found at the upper edge of the region, the region will be divided.

The points are now found as convexity defects in the same way as described for the tall regions. Instead of measuring the angle this method uses the y-coordinates of the points. The found point must be located on the upper edge of the region and have a y-value greater than both the convexity defect's start and end point to make it a minimum point.

As for splitting the tall regions, the algorithm will continue until no more regions are split.

Sort Regions

The final step is to sort the regions. After the regions have been split and the reflections have been removed, the remaining regions are now investigated before they are counted as a person. If a region's area does not match its distance to the camera it is removed. This could be a small region which is found in the foreground where persons typically would be larger. This step also calculates the person's position on the court, using the homography from the initialisation. This is done for the lowest middle pixel of the accepted regions, which will be the position on the floor.

4.3.4 Occupancy of the Court

A user's position will be given as a x,y coordinate with multiple decimals. In order to examine the occupancy a method must be found that preserves the position, but also mimics the size of a person. Therefore every found region is represented as a 3D gaussian distribution with a height of 1 and $\sigma = 5$, equivalent to a radius of 1 metre for 95 % of the volume. This is also roughly the radius of a person. An example for one frame can be seen in figure 4.4 where 8 persons have been found.

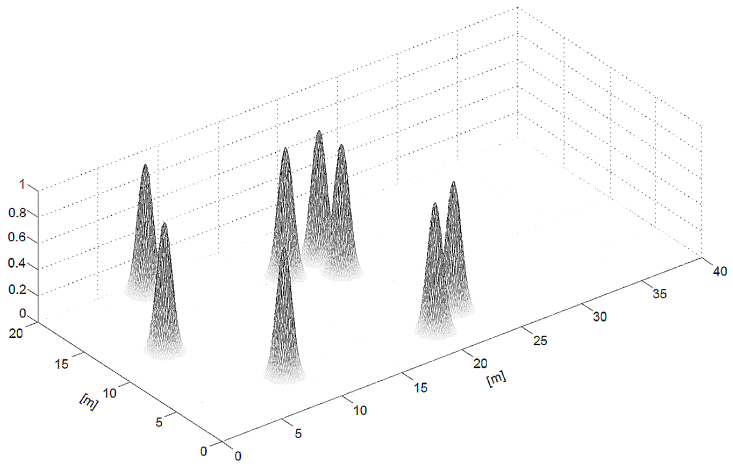


Fig. 4.4: A single frame where 8 persons have been found.

For longer periods these frames can be summed to show the occupancy of the court during e.g. an hour.

4.4 Results

4.4.1 Objective

As described in section 4.2 a very important parameter for the system is the stability in changing conditions and in changing set-ups. Therefore the system is tested in five different arenas, capturing more than three weeks altogether. For all tests in different arenas the same parameters of the system has been used. Only the initialisation of the system, described in section 4.3.2, depends on the arena. By measuring the entropy of a number of frames with and without people, the entropy threshold TH is chosen to be 4.1. The thermal camera used in the test is an AXIS Q1921-E, with a resolution of 384×288 pixels and a horizontal field-of-view of 55°.

4.4.2 Annotation of data

Capturing three weeks continuously with 30 fps gives a total of 54,432,000 frames, which would be nearly impossible to manually annotate. Therefore it is chosen to manually annotate 54,000 frames, resulting in 30 minutes of video. This will be used for calculating the precision of the system. But as this test does only evaluate the system during one specific activity an additional test will be conducted. A period of 36 consecutive hours will be sampled and manually annotated with 0.04 fps (1 frame per 25 seconds). This is covering two days with different sports activities and a night. Even though the frame rate here is low this will still give a good evaluation of the system and ensure that it is tested with both a varying number of people and different types of sports. The data from the full test period of more than three weeks will be evaluated by random checks against the videos.

4.4.3 30 minutes test

The results for the 30 minutes period are calculated as a mean number for every five minutes. The automatic results compared to the ground truth (manually annotated data) are shown in figure 4.5 with black and red.

Calculating the error of the automatic system, sampled with 30 fps, for each five minutes period gives an average error of 20.5 %. Comparing the green line to the black line shows that in 4 out of 6 periods the automatic results with different sample rates are nearly the same, while for the last two periods the difference is about 0.5 person. From this it is concluded that even with a sample rate decreased to 0.04 fps the results will still be reliable.

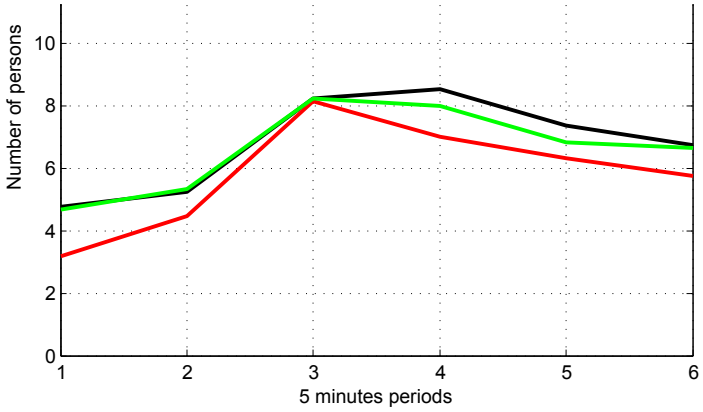


Fig. 4.5: Manual (red) and automatic (black) result for six five minutes periods, sampled with 30 frames per second, and automatic results (green) sampled with 0.04 frames per second.

4.4.4 Two days test

For the 36 hours, sampled with 1 frame per 25 seconds, a mean error is found for every five minutes, and stated as a mean error for each hour, since the activities in the arena are typically the same for at least an hour. This method is used for both the error measured in persons and per cent. The hours are then categorised by the maximum number of people, to investigate the relation between the error and the number of persons. See the results in table 4.1.

# persons	# hours	Mean error	Mean error (%)
0	12	0.0017	0.17 %
1-2	15	0.0428	7.35 %
7-15	9	0.5100	11.76 %

Table 4.1: Error categorised by the maximum number of people during the hour.

For the nine hours with maximum 7-15 persons on the court the error for each hour lies from 4-20 %, with an average of 11.76 % as stated in table 4.1. It is clear that the error for detecting empty arenas is very low, and the error increases with the number of persons. This will typically be due to occlusions. As mentioned in section 4.3.1 the occupancy level should be categorised to zero, some or many users, which means that the precise number of people is not critical for this application.

The 30 minutes test described in section 4.4.3 showed an error of 20.5 %, which is equivalent to the maximum error found during this two day test. The video of 30 minutes had a high activity level and a highly varying number of people on the court, with up to 14 people in each frame. Therefore it is also expected that it should have a higher error than the average videos.

4.4.5 Evaluation of positions

The calculated positions of the persons will be evaluated by visually comparing the manually marked positions with the automatic found positions. This is done for the 36 hours sampled with 1 frame per 25 seconds. An example of one hour showing a handball match can be seen in figure 4.6. The upper image shows the positions found by the system and the lower image shows the true position found manually for the same period. There are found more people manually than automatic during this hour, resulting in darker colours in the bottom image, but it is evident that there is a high correlation between the two images and the overall picture is the same. Note that the camera could only see the left half of the court.

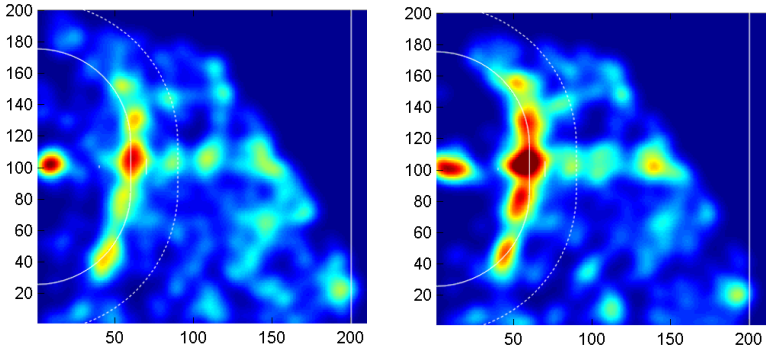


Fig. 4.6: Positions of users during a handball match. Left: Automatic. Right: Ground truth. Note that there are found more people manually than automatic, resulting in darker colours in the bottom image.

As mentioned in section 4.1 the position should be used to examine whether the entire court is being used or only part of it. Therefore the main point in evaluating the found positions is not to examine the position of each person, but to ensure that the overall picture of the occupancy during a booking is correct. This correlation between automatic and manually found positions is found to be very high for all 36 hours.

4.4.6 One week evaluation

The main objective for this system is to analyse the use of the sports arenas. Most sports arenas in Denmark have a booking system, where the local schools and sports clubs book their hours in the arena. To evaluate the use of the arenas the bookings should be considered. Seven consecutive days in one sports arena has been chosen, and the use is here measured as a mean number of persons per hour. The number is categorised as zero, some or many persons to describe the

level of occupancy. Table 4.1 showed that the precision of the system depends on the number of persons, the error increases when the number increases. As the error is very low for detecting empty arenas and few people on the court, the error of the exact number will not have a visible effect on the categorisation.

Finally the utilisation is compared to the booking as shown in figure 4.7. White areas are not booked, red areas are booked but never used, orange areas are booked and used by two or less persons in average, the green area are booked and used by more than two persons in average while the blue areas are used by more than two persons, but not booked. During this test a frame rate of 1 fps has been used.

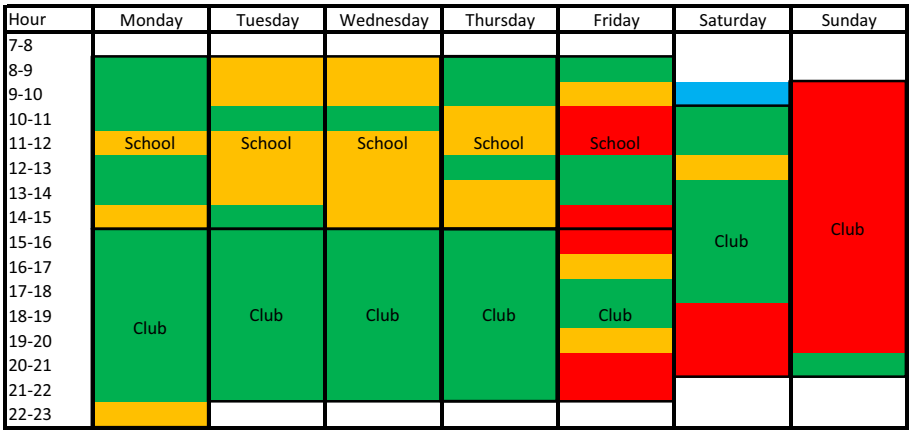


Fig. 4.7: Table of utilisation compared to the booking. White areas are not booked, red areas are booked but never used, orange areas are booked and used by two or less persons in average, green areas are booked and used by more than two persons in average and blue areas are used by more than two persons, but not booked.

Figure 4.7 indicates that during the measured seven days 21.2 % of the booked hours are not used, while 23.4 % are used by an average of two or less persons, which either means that the arena has only been used for a very short period of the hour, or there have been only one or two people at the court. One hour are used but not booked, which could also be a problem, depending on the policy for the administration of the arena.

4.5 Conclusion

This work presented an approach for automatic detection of persons using thermal cameras. For the intended application in sports arenas the privacy issue is important, therefore a thermal camera is chosen.

The system shows very satisfactory results, with only a short initialisation it works independently of the changing conditions in different arenas. The system can easily distinguish between an empty arena, few or many people.

The work will continue with further tests of the system and work on improving the segmentation of people. This could be by including temporal information or by using a more detailed human template for comparison with the found regions. For future work there are a lot of possibilities for developing new features, including analysis of the activity level, activity type and user type.

Acknowledgements

We would like to thank Aalborg municipality for support and for providing access to the sports arenas.

References

[1] M. Pilgaard, *Sport og Motion i Danskernes Hverdag (Sport and Exercise in the Everyday Life of Danish People)*. Idrættens Analyseinstitut, October 2009.

[2] S. Brixen, K. H. Larsen, J. V. Lindholm, K. F. Nielsen, and S. Riiskjær, *Strategi 2015: En Situationsanalyse (Strategy 2015: A Situation Analysis)*. DGI, 2010.

[3] C. J. Needham and R. D. Boyle, "Tracking multiple sports players through occlusion, congestion and scale," in *British Machine Vision Conference*, 2001, pp. 93–102.

[4] H. Saito, N. Inamoto, and S. Iwase, "Sports scene analysis and visualization from multiple-view video," in *Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on*, vol. 2, june 2004, pp. 1395 –1398 Vol.2.

[5] J. Xing, H. Ai, L. Liu, and S. Lao, "Multiple player tracking in sports video: A dual-mode two-way bayesian inference approach with progressive observation modeling," *Image Processing, IEEE Transactions on*, vol. 20, no. 6, pp. 1652 –1667, june 2011.

[6] T. Ko, "A survey on behavior analysis in video surveillance for homeland security applications," in *Applied Imagery Pattern Recognition Workshop, 2008. AIPR '08. 37th IEEE*, oct. 2008, pp. 1 –8.

[7] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 11, pp. 1473 –1488, nov. 2008.

[8] W. Wei and A. Yunxiao, "Vision-based human motion recognition: A survey," in *Intelligent Networks and Intelligent Systems, 2009. ICINIS '09. Second International Conference on*, nov. 2009, pp. 386 –389.

[9] T. B. Moeslund, A. Hilton, V. Krüger, and L. Sigal, *Visual Analysis of Humans - Looking at People*. Springer, 2011.

[10] W. K. Wong, P. N. Tan, C. K. Loo, and W. S. Lim, "An effective surveillance system using thermal camera," in *Signal Acquisition and Processing, 2009. ICSAP 2009. International Conference on*, april 2009, pp. 13 –17.

[11] W. K. Wong, Z. Y. Chew, C. K. Loo, and W. S. Lim, "An effective trespasser detection system using thermal camera," in *Computer Research and Development, 2010 Second International Conference on*, may 2010, pp. 702 –706.

[12] W. Wang, J. Zhang, and C. Shen, "Improved human detection and classification in thermal images," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, sept. 2010, pp. 2313 –2316.

[13] M. Bertozzi, A. Broggi, P. Grisleri, T. Graf, and M. Meinecke, "Pedestrian detection in infrared images," in *Intelligent Vehicles Symposium, 2003. Proceedings. IEEE*, june 2003, pp. 662 – 667.

[14] J. Davis and V. Sharma, "Robust detection of people in thermal imagery," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 4, aug. 2004, pp. 713 – 716 Vol.4.

[15] A. Criminisi, "Computing the plane to plane homography," 1997. [Online]. Available: http://www.robots.ox.ac.uk/~vgg/presentations/bmvc97/criminispaper/node3.html

[16] J. Kapur, P. Sahoo, and A. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Computer Vision, Graphics, and Image Processing*, vol. 29, no. 3, pp. 273 – 285, 1985.

[17] S. Suzuki and K. Abe, "Topological structural analysis of digitized binary images by border following," *Computer Vision, Graphics, and Image Processing*, vol. 30, no. 1, pp. 32 – 46, 1985.

[18] D. S. DST, "Tabel 44: De værnepligtiges højde (conscripts' height in 2006)," 2006. [Online]. Available: http://www.dst.dk/aarbogstabel/44

Chapter 5

Long-term Occupancy Analysis using Graph-Based Optimisation in Thermal Imagery

Rikke Gade, Anders Jørgensen and Thomas B. Moeslund

The paper has been published in the
Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3698–3705, June 2013.

© 2013 IEEE
The layout has been revised.

Abstract

This paper presents a robust occupancy analysis system for thermal imaging. Reliable detection of people is very hard in crowded scenes, due to occlusions and segmentation problems. We therefore propose a framework that optimises the occupancy analysis over long periods by including information on the transition in occupancy, when people enter or leave the monitored area. In stable periods, with no activity close to the borders, people are detected and counted which contributes to a weighted histogram. When activity close to the border is detected, local tracking is applied in order to identify a crossing. After a full sequence, the number of people during all periods are estimated using a probabilistic graph search optimisation. The system is tested on a total of 51,000 frames, captured in sports arenas. The mean error for a 30-minute period containing 3-13 people is 4.44 %, which is a half of the error percentage optained by detection only, and better than the results of comparable work. The framework is also tested on a public available dataset from an outdoor scene, which proves the generality of the method.

5.1 Introduction

Measuring the occupancy maps from people has become an essential step towards an intelligent and efficient society [1, 2]. A well-known example of this is that the whereabouts of people in shopping malls provides valuable information for the managers. The same goes for sports arenas. These facilities are in high demand, but very expensive to build, so focus of the political systems has shifted towards optimising the use of the existing arenas. The first step in this analysis is to monitor the occupancy of such facilities. As this analysis should run for several weeks in each arena, manual observations would be expensive and cumbersome, and an automatic system based on computer vision is therefore suggested. While RGB-based systems are normally used in previous research in sports analysis [3–6], a general public acceptance of more permanent installations in such facilities are harder to come by due to privacy issues. We therefore apply thermal imagery, which captures the infrared radiation instead of visible light, and creates an image whose pixel values represent temperature. People can not be identified in thermal images, thereby eliminating the privacy issues. A positive side effect of thermal imaging is that detection can often be reduced to a trivial task. However, thermal imaging also introduces new problems, as people are often fragmented into small parts, and reflections can be seen in the floor. Moreover, the challenges of occlusions remain in thermal images, see figure 5.1.

The contribution of this work is a reliable method for occupancy analysis in thermal video. The method does not assume a perfect detection in each frame, but handles the detection challenges by including temporal information. The main focus is not short lab sequences, but rather long, real-life sequences. Here

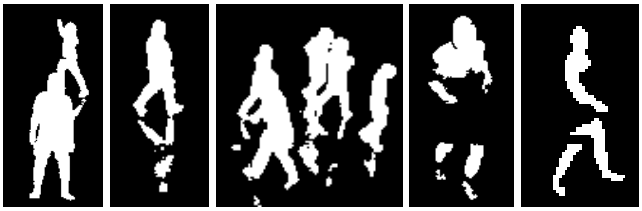


Fig. 5.1: Examples of the challenges for detection of people.

we use data from sports arenas, which are very challenging, due to the natural physical interaction in sport.
The main idea is to split the video sequences into two types of periods. The first type is the stable periods, where no people exit or enter the court. In these periods, the number of people on the court must be the same, which in turn introduces a constraint on the problem. The second type defines unstable periods, where the occupancy is likely to change. Combining these two types of information to model the periods and transitions between them provides a unified framework to optimise over a long period of time.

5.1.1 Thermal radiation

Thermal imaging is still a relatively new modality in computer vision applications, and the theory behind it is relatively unknown in the computer vision society. This section will therefore provide information on the physical foundation of thermal radiation and cameras.
All objects with a temperature above the absolute zero emit infrared radiation, mainly in the mid-wavelength infrared spectrum (MWIR, 3-5 μm) and long-wavelength infrared spectrum (LWIR, 8-15 μm). This is often referred to as thermal radiation. The intensity of the radiation from an object with temperature T is described by Planck's Law as a function of the wavelength λ:

$$I(\lambda, T) = \frac{2\pi hc^2}{\lambda^5 \left(e^{hc/\lambda k_B T} - 1\right)} \tag{5.1}$$

where h is Planck's constant ($6.626 \times 10^{-34} Js$), c the speed of light ($299,792,458 m/s$) and k_B Boltzmann's constant ($1.3806503 \times 10^{-23} J/K$). From this expression, it can be seen that the intensity peak shifts to shorter wavelenghts as the temperature increases. For extremely hot objects, the radiation extends into the visible spectrum.
The thermal radiation originates from energy in the molecules of an object. The energy can be expressed as a sum of four contributions [7]:

$$E = E_{electronic} + E_{vibration} + E_{rotation} + E_{translation} \tag{5.2}$$

Only the energy caused by translation, rotation and vibration in a molecule contributes to the temperature of an object.

It is well-known from quantum physics, that visible light consists of photons that causes electron transitions when they are absorbed or emitted from a molecule. The same principle applies to infrared light, with the difference that the photons contain less energy and cause transitions in the vibrational and rotational energy levels instead. The electromagnetic radiation can be absorbed or emitted by the molecule, then the incident radiation causes the molecule to rise to an excited energy state, and when it falls back to ground state a photon is released. Only photons with specific energies, equal to the difference between two energy states, can be absorbed and emitted.
If more radiation is absorbed than emitted, the temperature of the molecule will rise until equilibrium is re-established. Likewise, the temperature will fall if more radiation is emitted than absorbed, until equilibrium is re-established.

5.1.2 Thermal cameras

Generally two types of detectors exist for thermal cameras: photon detectors and thermal detectors. Photon detectors convert the absorbed electromagnetic radiation directly into a change of the electronic energy distribution in a semiconductor by the change of the free charge carrier concentration. This type of detector typically works in the MWIR spectrum, where the thermal contrast is high, making it very sensitive to small differences in the scene temperature. The main drawback is the need for cooling of the detector, making it more expensive and with a higher need for maintenance. The thermal detector converts the absorbed electromagnetic radiation into thermal energy causing a rise in the detector temperature. Then, the electrical output of the thermal sensor is produced by a corresponding change in some physical property of material, e.g., the temperature-dependent electrical resistance in a bolometer. This type of detector measures radiation in the LWIR spectrum. They are uncooled and have been developed with two different types of sensors: ferroelectric detectors and microbolometers, where today the microbolometer has shown to have more advantages.

5.1.3 Related work

Detection of people is the first step in many applications, e.g. surveillance, tracking, or activity analysis. General purpose detection systems should be robust and independent of the environment. The thermal cameras can here often be a better choice than a normal visual camera.
The methods applied to thermal imaging span from simple thresholding and shape analysis [8–12] to more complex, but well-known methods such as HOG and SVM [13–17] as well as contour analysis [18–21]. Using simple methods allows for fast real-time processing, and combined with the illumination independency, the thermal sensor is very well suited for detecting humans in real-life applications.

An obvious application area for thermal imaging is pedestrian detection systems for vehicles, due to the cameras' ability to "see" during the night. These systems are being developed both as assistance for drivers in low visibility, and as a navigation tool for the future automatic vehicles. One of the car-based detection systems is proposed in [22], where they present a tracking system for pedestrians. It works well with both still and moving vehicles, but some problems still remain when a pedestrian enters the scene running. [23] proposes a shape-independent pedestrian detection method. Using a thermal sensor with low spatial resolution, [24] builds a robust pedestrian detector by combining three different methods. [25] also proposes a low resolution system for pedestrian detection from vehicles. [26] proposes a pedestrian detection system that detects people based on their temperature and dimensions, and tracks them using a Kalman filter. In [27] a stereo-vision system has been tested, detecting warm areas and classifying if they are humans, based on distance estimation, size, aspect ratio, and head shape localisation.
A more general interest in pedestrian detection based on thermal imaging can also be seen in surveillance or for analysis of pedestrian flow in cities. A general purpose pedestrian detection system is proposed in [28]. The foreground is separated from the background, after that shape cues are used to eliminate non-pedestrian objects and appearance cues help to locate the exact position of pedestrians. A tracking algorithm is also implemented. [29] uses probabilistic template models of four different poses for detection. [30] also uses probabilistic template models, here they use three models representing different scales. [31] uses a statistical approach for head detection as the first step in the pedestrian detection.
The previously described methods use thermal sensors only. Combining different types of sensors could, however, eliminate some of the disadvantages from both sensors. Examples of systems combining thermal and RGB cameras are given by Davis et al. [19, 32] and Leykin et al. [33, 34]. Other sensors like laser scanners and near-infrared cameras, have also been combined with thermal sensors [35, 36].
Due to privacy issues, this work will concentrate on thermal cameras only. We will also take advantage of the easy foreground segmentation, but as shown in figure 5.1, challenges still remain. As opposed to most existing work, it will be tested on long sequences of real data with high complexity.

5.2 Approach

As described in the introduction, precisely counting people in single frames can be a nearly impossible task, due to occlusions and segmentation errors. Therefore, it is suggested to include temporal information, and estimate the occupancy over longer periods. The idea is to automatically split a video sequence into stable periods, with no activities near the border of the court, and transition periods with activity near the border. During the stable periods, the

detected number of people in each frame contributes to a distribution of observations for that period. For the transition periods, local tracking of the blobs in the border area is applied, in order to estimate the likelihood of crossings. The two types of data and their uncertainties are combined in a graph, where the nodes represent the number of people, and the edges represent the change in number between two periods. A dynamic programming approach is applied to find the optimal path of the graph.
The remaining part of section 5.2 describes the details of the people detection and the monitoring of transitions. In section 5.3 the graph optimisation is described, and in section 5.4 the system is evaluated. The conclusion is found in section 5.5.

5.2.1 People detection

The first step towards detecting people is to separate foreground from background. Using thermal imagery in an indoor environment simplifies this task, as the surrounding temperature is normally stable and colder than the human temperature. There can, however, be observed warm spots, e.g, from heaters, hot water pipes, and doors or windows heated by the sun. A background subtraction method is used to remove static objects from the foreground. Since the image depicts the temperature of an indoor scene, it can be assumed that only slow changes will occur in the background. Therefore, the background image simply consists of the average of the previous n frames, but only pixels that are classified as background will contribute to the new background estimate.
Even though the foreground is now found, pixel noise should be removed. Moreover, due to the camera having automatic gain adjustment, the level of pixel values can suddenly change, without any temperature changes in the scene. To overcome these challenges, an automatic threshold method based on maximum entropy is used to calculate the threshold value for each frame [37]. From this point the image is binary, and all blobs found are considered potential persons. The next part, section 5.2.2 and 5.2.3 will deal with the splitting and sorting of blobs into single persons.

5.2.2 Groupings

Since a side-view of the scene is obtained, see section 5.4.1, it is necessary to be able to handle occlusions. Generally, two types of occlusions are seen: people standing behind each other, seen from the camera's point of view ("tall blobs") and people standing close together in a group ("wide blobs").

Split tall blobs

In order to split people that form one blob by standing behind each other, it must be detected when the blob is too tall to contain only one person. We here adapt the method from [9]. If the blob has a pixel height that corresponds

to more than a maximum height at the given position, see section 5.4.1, the algorithm should try to split the blob horizontally. The point to split from is found by analysing the convex hull and finding the convexity defects of the blob. Of all the defect points, the point with the largest depth and a given maximum absolute gradient should be selected, meaning that only defects coming from the side will be considered, discarding e.g. a point between the legs. See examples in figure 5.2.

Split wide blobs

People standing close to each other, e.g., in a group, will often be found as one large blob. To identify which blobs contain more than one person, the height/width ratio and the perimeter are considered, as done in [9]. If the criteria are satisfied, the algorithm should try to split the blob. For this type of occlusion, it is often possible to see the head of each person, and split the blob based on the head positions. Since the head is more narrow than the body, people can be separated by splitting vertically from the minimum points of the upper edge of a blob. These points can be found by analysing the convex hull and finding the convexity defects of the blob. See examples in figure 5.2.

Fig. 5.2: Examples of wide and tall blobs that have been split.

5.2.3 Sorting people candidates

In addition to occlusions, other problems like reflections from people in the floor, or one person split into many blobs can be observed. This means that blobs can not always be mapped into individual people. In order to solve these challenges, the idea of generating a probabilistic occupancy map [38, 39] is adapted to find the probability that a person is observed at a given location. The original ideas were applied for multi-camera tracking, where it is possible to observe the 3D location of the scene. For this work, part of the idea is adapted to work on binary objects, captured from a single view. The algorithm will take all the bottom points of the blobs as person location candidates, and calculate the probability for each of them being a true position. A rectangle is generated from each candidate point, with a height corresponding to a given average height of people and the width being one third of the height. Two parameters are used for evaluating the probability of the rectangle containing a person:

the ratio of white pixels inside the rectangle and the ratio of the rectangle perimeter that is white. Figure 5.3 shows two histograms of the ratio of white pixels inside the rectangles for true candidates (blue) and false candidates (red). The histograms are built from manual annotation of 340 positive samples and 250 negative samples.

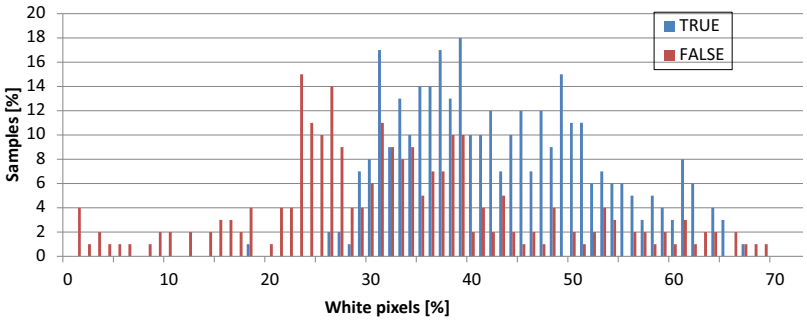


Fig. 5.3: Histograms of the percentage of white pixels in each candidate rectangle. The blue histogram is for true candidates and red is for false candidates. No samples are found above 70 %.

From figure 5.3 it is seen that only 1 % of the true candidates have a white ratio less than 25 %, while a large part of the false candidates are found here, and no true candidates are above 70 %.
For the rectangle perimeter it is found that the lower the ratio of the rectangle perimeter that is white, the better is the fit of the rectangle to the person. The weighting of a person, $w_p(i)$, is described in equation 5.3 from the ratio of white pixels in the rectangle, r_r, and the ratio of white pixels on the perimeter, r_p:

$$w_p(i) = \begin{cases} 0, & \text{if} r_p > 50\% \parallel r_r < 20\% \\ 0.8, & \text{if} r_r > 70\% \\ 0.9, & \text{if} r_r < 30\% \parallel r_r > 60\% \\ 1, & \text{otherwise} \end{cases} \tag{5.3}$$

Candidates with $w_p(i) = 0$ are deleted.
There are still a lot of false candidates that will not be affected by these criteria. Many of them contain part of a person, and overlap in the image with a true candidate. Due to the possibility of several candidates belonging to the same person, the overlapping rectangles must be considered. By tests from different locations and different camera placements, it is found that if two rectangles overlap by more than 60 %, they probably originate from the same person, or from reflections of that person. As only one position should be accepted per person, only one of the overlapping rectangles should be chosen. Due to low resolution images compared to the scene depth, cluttered scenes, and no restrictions on the posture of a person, the feet of a person can not be recognised

from the blobs. Furthermore, due to the possibility of reflections below a person in the image, it can not be assumed that the feet are the lowest point of the overlapping candidates. Instead, the best candidate will be selected on the highest ratio of white pixels, as it is seen from figure 5.3, that the probability of false candidates are lower here.

5.2.4 Identification of people entering and leaving

During the periods with activities detected at the border of the court, it is very likely that a change will happen. For these periods, the people near the border are monitored in order to detect crossings. The people are detected as described in section 5.2.3, but will not be counted during these unstable periods. Instead, the position of each person near the border is tracked, and if the border is crossed, it is registered along with the direction. Until a new stable period is observed, the number of people entering or leaving the court will contribute to the total transition in number.

5.3 Graph search optimisation

Two types of data exist now, the number of detected persons during the stable periods, and the number of entering or leaving persons during periods with activity at the border. The last step is now to combine these estimates in a graph for the total observed period and estimate the most probable number during all periods. The graph will consist of nodes, representing the number of people in the stable periods and edges, representing the change in number between two periods. Figure 5.4 is a simple example of a graph with three stable periods. Edges exist between all nodes in two consecutive periods, but to simplify the illustration they are not drawn.

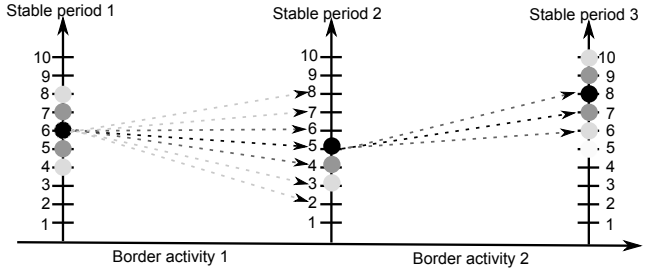


Fig. 5.4: Example of a simple graph. Dark nodes and edges have the highest weight. Edges exist between all nodes in two consecutive periods, but to simplify the illustration they are not drawn.

A dynamic programming approach is taken to calculate the optimal path. The problem is solved by a version of Dijkstra's Algorithm modified to calculate

the path with the highest votes, instead of the traditional minimum cost. The probabilistic weighting of nodes and edges will be described in the next section.

5.3.1 Weighting

Each node and edge must be weighted in order to calculate the best path. We define the weights as positive, meaning that a higher weight is a better path. As described, each node in the graph represents a possible number of persons in a given period. The weights for the nodes will be distributed according to the weighted histogram of the number of detected people in all frames during the stable period. The histogram is constructed from the detected people in each frame, with a weight describing the probability of each detection being true, and a weight describing the uncertainty of the frame, caused by occlusions and clutter. Each frame counting is weighted like this:

$$w_f = \alpha \cdot \prod_{i=1}^{n} w_p(i) + \beta \cdot w_s \tag{5.4}$$

where n is the number of people, $w_p(i)$ is the probability of people i being a true detection (see equation 5.3), and w_s is a weight that decreases with the number of splits performed (described in section 5.2.2), indicating how cluttered the scene is. α and β are the weighting of each part and should sum to one. The observed number in a frame will be added to the histogram with the weight w_f, and after a stable period has ended, the histogram will be scaled to an accumulated sum of 1. The circles in figure 5.4 illustrate the weighted histogram for each period.

The weighting of edges depends on the total number of crossings during the period of border activity, as well as the weighting of the individual people crossing the border. The probability of change x in number of people ($+n$ for people entering the court and $-n$ for people leaving) is modelled as a Gaussian distribution, with the mean value μ being the calculated number, and the variance σ proportional to the total number of crossings. The probability is described as $w_b(x)$:

$$w_b(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)} \times w_p \tag{5.5}$$

$$w_p = \frac{1}{m} \sum_{i=1}^{m} w_p(i) \tag{5.6}$$

where m is the number of people crossing the border. Each dashed line in figure 5.4 illustrates the edges weighted with $w_b(x)$. In the example the variance σ is high for the first period of border activity and low for the second period of border activity.

5.4 Experimental results

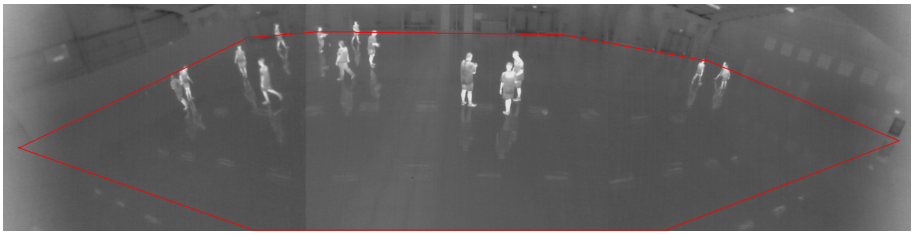


Fig. 5.5: Example of the thermal image, with the outline of the court drawn as a red line.

Comparing our results with others is difficult, because as far as we know, only [9] has focused on occupancy analysis of thermal video. We therefore compare our work to related work based on RGB cameras. Moreover, no public datasets with long thermal videos containing more than a few people exist. We therefore capture a new dataset, that will be available for download after publication [1]. The data contained in this video is from six different arenas, in order to be able test the robustness of the algorithms in different environments and set-ups. Several different activities are captured as well as both children and adults. We test on a 5-minute sequence from each of the five arenas for the evaluation of the detection algorithm and the tracking algorithm for the border areas. The full system with graph search optimisation should benefit from a longer video sequence, and will therefore be tested on a 30 minute video from a sixth arena. Thereby, the system has been tested on a total of 51,000 frames, which are manually annotated to provide ground truth. This data contains between 3-16 people in each frame. The processing time is approximately 0.125 seconds per frame on an Intel Core 2 Duo 3 GHz CPU, without any optimisation of the software.
To prove the generality of our framework, a final test has also been conducted on a public dataset of a totally different scenario, which is an outdoor scene from the OSU Color/Thermal database [32]. This test will be described in section 5.4.5.
The remaining part of this section will describe the calibration and initialisation needed for the system, before results for each test are presented.

5.4.1 Camera calibration and initialisation

Installing a camera in the ceiling above most courts is very cumbersome and expensive and therefore not realistic in general. Therefore, it must be installed on one of the walls or rafters around the court. A standard arena has a court of 40×20 metres, corresponding to a handball field, indoor soccer field, etc. As the lenses of commercial thermal cameras today have a maximum field-of-view

[1] Available for download at www.vap.aau.dk

of around 60°, more than one camera must be used to cover the entire court. The camera set-up used in this work consists of three thermal cameras placed at the same location, and adjusted to have adjacent fields-of-view. Each camera is of the type Axis Q1922, which uses an uncooled microbolometer for detection. The resolution is 640 × 480 pixels per camera and the horizontal field-of-view is 57° per camera. To make the system invariant to the cameras' set-up, the images are stitched together before processing. This requires the cameras to be perfectly aligned and undistorted in order to secure smooth "crossings" between two cameras. Calibration of thermal cameras is not a trivial task, as they can not see the contrast differences of a typical chessboard used in most applications. Therefore, a special calibration board is made with 5x4 small incandescent light bulbs. With this it is possible to adapt the traditional method to estimate the intrinsic parameters of the cameras. The cameras are manually aligned horizontally so that their pitch and roll are the same. This mimics the well-known panorama effect, but with three cameras capturing at the same time. An example of the resulting image is shown in figure 5.5. When the cameras are put up in an arena, an initialisation is made. This consists of finding the mapping between image and world coordinates, as well as finding the correlation between peoples' real height and their height in the images, corresponding to their distance to the camera.
As the cameras are fixed relative to each other and then tilted downwards when recording in arenas, the result is that people in the image are more tilted the further they get from the image centre along the x-axis. This means that a person's pixel height can not always be measured vertically in the image. Therefore, the calibration must include both the height and the angle of a person standing upright at predefined positions on the court. For this work we used points on a grid of 5 × 5 metres on the court resulting in 45 different calibration images. In each image the world coordinates, image coordinates, pixel height and angle are learned as well as the person's real height in metres. The four corners are used to calculate a homography for each square, making it possible to map image coordinates to world coordinates. Using interpolation, an angle and maximum height are calculated for each position.

5.4.2 Detection of people

The first test evaluates the detection algorithm described in section 5.2.1. The number of detected people is registered as well as the manually counted number. This is done for 5 videos of 5 minutes each, captured with 10 fps, altogether 15,000 frames.
The mean error for each video is found to be between 8.5 % and 22.0 %. The errors are independent of the arena and seems primarily to depend on the level of occlusions seen in the scene. Periods with large groupings have a higher detection error than periods with people separated from each other. This is also expected, as the detection algorithm works on each frame independently, and people that are fully or mostly occluded can not be detected. Apart from

the initialisation described in section 5.4.1, nothing has been done to fit the system to the specific arena, and it is concluded that it is independent of the arena.

5.4.3 Transition recognition

For the five videos of five minutes, it is registered each time a person crosses a specified border in order to evaluate the tracking algorithm. A total of 154 crossings are detected manually, and 168 crossings are detected automatically. 108 of the crossing are detected at the exact time, which is considered within $\pm$ 2 frames of the manual detection. Most of the false crossings detected are compensated with a crossing in the opposite direction within a few frames. These will therefore not affect the global estimation of the number.

5.4.4 Full system test

The full system is tested on a 30 minute video, captured with 20 fps. Calculating the error for each frame gives an average error of 0.38 persons, corresponding to 4.44 %. For comparison, the result using detection only is also found, the error here is twice as high, 8.87 %. The number of detections is very unstable, and could suggest to do a simple low pass filtering, to overcome what looks like high frequency noise in the measurements. Low pass filtering the detection data reduces the error to 7.70 %. This indicates that a simple filtering of the data will not reduce the error as efficiently as the graph optimisation method. In table 5.1 our results are compared to related work, based on both thermal and RGB images.

	Reported error
Gade et al. [9]	7.35-11.76 %
Rabaud and Belongie [40] *	6.3-10 %
Hou and Pang [41] *	10 %
Celik et al [42] */**	8 - 14 %
Our method	**4.44** %

Table 5.1: Reported error percentage from related work compared to our result. * uses RGB images. ** calculates the error as percentage of frames with an error larger than one person.

5.4.5 Test on OSU dataset

To show the generality of our framework, we tested the system on the thermal video from the OSU Color-Thermal database [32], which is dataset three from the OTCBVS Benchmark Dataset Collection. We used sequences 4, 5 and 6, which are videos of approximately one minute each. They contain between zero and four people in each frame. Due to the low number of people in this dataset,

instead of error we calculated the precision, being the number of frames with the correct number of people estimated. The results are presented in table 5.2 and compared to the results of detection alone, as well as the results of [43], which were provided with the dataset. However, it should be noted that the results of [43] are obtained by fusing the thermal and visible modalities and are intended for people tracking.

	Seq. 4	Seq. 5	Seq. 6
Detection only	86.72 %	83.11 %	77.72 %
Leykin et al. [43]	85.52 %	88.77 %	64.89 %
Our full method	**87.12** %	**93.70** %	**87.89** %

Table 5.2: Counting precision on the OSU dataset.

It is seen that the results of our full method are better than both the results from [43] and from detection alone.

5.5 Conclusion

In this work we have presented a unified framework for occupancy analysis. This method includes temporal information in the estimation by measuring the transition in numbers, and using that together with the detection of people in the global optimisation. The application of this work is the analysis of a given facility over days, weeks or even months. The need for real-time analysis is minor, and offline processing therefore allows for a more global approach. The main focus was on sports arenas, but we also proved that it works well in a general outdoor scene. We have shown that including the transition information improves the precision significantly, compared to using detection alone; even if the detection results are filtered afterwards. The mean error for the 30-minute test is 4.44 %, compared to 8.87 % if only the detection method was used.
The occupancy analysis is the foundation in many applications and can be continued to further activity analysis.

References

[1] R. Kitchin and M. Dodge, *Code/Space: Software and Everyday Life.* MIT Press, 2011.

[2] E. Poulsen, H. Andersen, R. Gade, O. Jensen, and T. Moeslund, "Using human motion intensity as input for urban design," in *Constructing Ambient Intelligence*, 2012.

[3] R. M. Barros, R. P. Menezes, T. G. Russomanno, M. S. Misuta, B. C. Brandao, P. J. Figueroa, N. J. Leite, and S. K. Goldenstein, "Measuring

handball players trajectories using an automatically trained boosting algorithm," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 14, no. 1, pp. 53 – 63, 2011.

[4] S. Kopf, B. Guthier, D. Farin, and J. Han, "Analysis and retargeting of ball sports video," in *IEEE Workshop on Applications of Computer Vision*, jan. 2011.

[5] E. F. de Morais, S. Goldenstein, and A. Rocha, "Automatic localization of indoor soccer players from multiple cameras," in *Proceedings of the International Conference on Computer Vision Theory and Applications*, feb. 2012.

[6] J. Xing, H. Ai, L. Liu, and S. Lao, "Multiple player tracking in sports video: A dual-mode two-way bayesian inference approach with progressive observation modeling," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1652 –1667, june 2011.

[7] R. A. Serway and J. W. Jewett, *Physics for Scientists and Engineers with Modern Physics*, 6th ed. Brooks/Cole—Thomson Learning, 2004.

[8] T. T. Zin, H. Takahashi, and H. Hama, "Robust person detection using far infrared camera for image fusion," in *Second International Conference on Innovative Computing, Information and Control*, 2007.

[9] R. Gade, A. Jørgensen, and T. B. Moeslund, "Occupancy analysis of sports arenas using thermal imaging," in *Proceedings of the International Conference on Computer Vision and Applications*, vol. 2, feb. 2012, pp. 277–283.

[10] W. K. Wong, Z. Y. Chew, C. K. Loo, and W. S. Lim, "An effective trespasser detection system using thermal camera," in *Second International Conference on Computer Research and Development*, 2010.

[11] A. Fernández-Caballero, J. C. Castillo, J. Serrano-Cuerda, and S. Maldonado-Bascón, "Real-time human segmentation in infrared videos," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2577 – 2584, 2011.

[12] C. Dai, Y. Zheng, and X. Li, "Layered representation for pedestrian detection and tracking in infrared imagery," in *CVPR Workshops*, june 2005.

[13] L. Zhang, B. Wu, and R. Nevatia, "Pedestrian detection in infrared images based on local shape features," in *CVPR*, 2007.

[14] F. Suard, A. Rakotomamonjy, A. Bensrhair, and A. Broggi, "Pedestrian detection using infrared images and histograms of oriented gradients," in *IEEE Intelligent Vehicles Symposium*, 2006.

[15] F. Xu, X. Liu, and K. Fujimura, "Pedestrian detection and tracking with night vision," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 1, pp. 63 – 71, march 2005.

[16] D. Olmeda, A. de la Escalera, and J. Armingol, "Contrast invariant features for human detection in far infrared images," in *IEEE Intelligent Vehicles Symposium*, 2012.

[17] W. Li, D. Zheng, T. Zhao, and M. Yang, "An effective approach to pedestrian detection in thermal imagery," in *Eighth International Conference on Natural Computation*, 2012.

[18] J. W. Davis and V. Sharma, "Robust detection of people in thermal imagery," in *ICPR*, 2004.

[19] J. W. Davis and M. A. Keck, "A two-stage template approach to person detection in thermal imagery," in *Seventh IEEE Workshops on Application of Computer Vision*, 2005.

[20] Z. Li, J. Zhang, Q. Wu, and G. Geers, "Feature enhancement using gradient salience on thermal image," in *International Conference on Digital Image Computing: Techniques and Applications*, 2010.

[21] W. Wang, J. Zhang, and C. Shen, "Improved human detection and classification in thermal images," in *17th IEEE International Conference on Image Processing*, 2010.

[22] E. Binelli, A. Broggi, A. Fascioli, S. Ghidoni, P. Grisleri, T. Graf, and M. Meinecke, "A modular tracking system for far infrared pedestrian recognition," in *IEEE Intelligent Vehicles Symposium*, june 2005.

[23] Y. Fang, K. Yamada, Y. Ninomiya, B. K. P. Horn, and I. Masaki, "A shape-independent method for pedestrian detection with far-infrared images," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 6, pp. 1679 – 1697, nov. 2004.

[24] M. Mahlisch, M. Oberlander, O. Lohlein, D. Gavrila, and W. Ritter, "A multiple detector approach to low-resolution FIR pedestrian recognition," in *IEEE Intelligent Vehicles Symposium*, 2005.

[25] J.-E. Kallhammer, D. Eriksson, G. Granlund, M. Felsberg, A. Moe, B. Johansson, J. Wiklund, and P.-E. Forssen, "Near Zone Pedestrian Detection using a Low-Resolution FIR Sensor," in *IEEE Intelligent Vehicles Symposium*, 2007.

[26] D. Olmeda, A. de la Escalera, and J. M. Armingol, "Detection and tracking of pedestrians in infrared images," in *Int'l Conference on Signals, Circuits and Systems*, 2009.

[27] M. Bertozzi, A. Broggi, C. Caraffi, M. D. Rose, M. Felisa, and G. Vezzoni, "Pedestrian detection by means of far-infrared stereo vision," *CVIU*, vol. 106, no. 2–3, pp. 194 – 204, 2007.

[28] C. Dai, Y. Zheng, and X. Li, "Pedestrian detection and tracking in infrared imagery using shape and appearance," *CVIU*, vol. 106, no. 2-3, pp. 288–299, May 2007.

[29] M. Bertozzi, A. Broggi, C. H. Gomez, R. I. Fedriga, G. Vezzoni, and M. Del Rose, "Pedestrian detection in far infrared images based on the use of probabilistic templates," in *IEEE Intelligent Vehicles Symposium*, june 2007.

[30] H. Nanda and L. Davis, "Probabilistic template based pedestrian detection in infrared videos," in *IEEE Intelligent Vehicle Symposium*, 2002.

[31] U. Meis, M. Oberlander, and W. Ritter, "Reinforcing the reliability of pedestrian detection in far-infrared sensing," in *IEEE Intelligent Vehicles Symposium*, 2004.

[32] J. W. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *CVIU*, vol. 106, no. 2–3, pp. 162 – 182, 2007.

[33] A. Leykin and R. Hammoud, "Robust multi-pedestrian tracking in thermal-visible surveillance videos," in *CVPR Workshop*, 2006.

[34] ——, "Pedestrian tracking by fusion of thermal-visible surveillance videos," *Machine Vision and Applications*, vol. 21, pp. 587–595, 2010.

[35] B. Fardi, U. Schuenert, and G. Wanielik, "Shape and motion-based pedestrian detection in infrared images: a multi sensor approach," in *IEEE Intelligent Vehicles Symposium*, 2005.

[36] R. Schweiger, S. Franz, O. Lohlein, W. Ritter, J.-E. Kallhammer, J. Franks, and T. Krekels, "Sensor fusion to enable next generation low cost night vision systems," *Optical Sensing and Detection*, vol. 7726, no. 1, 2010.

[37] J. Kapur, P. Sahoo, and A. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Computer Vision, Graphics, and Image Processing*, vol. 29, no. 3, pp. 273 – 285, 1985.

[38] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *PAMI*, vol. 30, no. 2, pp. 267 –282, feb. 2008.

[39] Y. Cho, Y. Choi, S. Bae, S. Lim, and H. Yang, "Multi-camera occupancy reasoning with a height probability map for efficient shape modeling," in *16th International Conference on Virtual Systems and Multimedia*, oct. 2010.

[40] V. Rabaud and S. Belongie, "Counting crowded moving objects," in *CVPR*, june 2006.

[41] Y.-L. Hou and G. Pang, "People counting and human detection in a challenging situation," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 41, no. 1, pp. 24 –33, jan. 2011.

[42] H. Celik, A. Hanjalic, and E. Hendriks, "Towards a robust solution to people counting," in *IEEE International Conference on Image Processing*, oct. 2006, pp. 2401 –2404.

[43] A. Leykin, Y. Ran, and R. Hammoud, "Thermal-visible video fusion for moving target tracking and pedestrian classification," in *CVPR*, 2007.

Part III

Activity recognition

After estimating the occupancy of a sports facility the questions "Are there anybody there?" and "How many are they?" can now be answered. For the next step towards a thorough understanding of how the facility is used, we ask the question "What are they doing?".

Indoor sports facilities belonging to schools or municipalities host a wide variety of activities for all age groups. Users include schools and local clubs from introductory level to semi-professionals. With a high number of different users in these arenas, the knowledge about their activities is often sparse, and manual observation or labelling of video data is very time consuming and expensive. In this work we take the first step towards a full analysis of the activities by automatically recognising a number of defined sports types observed in the arena.

The two chapters in this part of the thesis present two different methods for classification of sports types, which were originally published in a book chapter and a workshop paper:

Rikke Gade and Thomas B. Moeslund, “Classification of Sports Types using Thermal Imagery,” *Computer Vision in Sports*, Springer, January 2015.

Rikke Gade and Thomas B. Moeslund, “Classification of Sports Types from Tracklets,” *KDD workshop on Large-Scale Sports Analytics*, August 2014.

Chapter 6

Classification of Sports Types using Thermal Imagery

Rikke Gade and Thomas B. Moeslund

The paper is to be published in
Computer Vision in Sports, January 2015.

© 2015 Springer
The layout has been revised.

Abstract

In this chapter we propose a method for automatic classification of five different sports types. The approach is based only on occupancy heatmaps produced from position data and is very robust to detection noise. To overcome any privacy issues when capturing video in public sports arenas we use thermal imaging only. This image modality also facilitates an easier detection of humans, the detection algorithm is based on automatic thresholding of the image. After a few occlusion handling procedures, the positions of people on the court are calculated using homography. Heatmaps are produced by summarising Gaussian distributions representing people over 10-minute periods. Before classification the heatmaps are projected to a low-dimensional discriminative space using the principle of Fisherfaces. We test our approach on two weeks of video and get very promising results with a correct classification of 89.64 %. In addition, we get correct classification on a publicly available handball dataset.

6.1 Introduction

In most societies, sport is highly supported by both governments and private foundations, as physical activity is considered a good way to obtain better health among the general population. The amount of money invested in sports facilities alone every year is huge, not only for new constructions, but also for maintenance of existing facilities. In order to know how the money is best spent, it is important to thoroughly analyse the use of the facilities. Such analyses should include information on how many people are present at any times, where they are, and what they are doing. The first two issues are the subjects of two recent papers by Gade et al. [1, 2].

In this chapter we will present a new method for automatic classification of sports types. We focus on the activities observed in public indoor sports arenas. In these arenas many types of activities take place, from physical education in schools, to elite training of a particular sport. For the administrators as well as the financial supporters of the arenas it is important to gain knowledge of how the arenas are used, and use that information for future planning and decisions on the lay-out of new arenas. In a future perspective it could also be of great value for the coaches and managers of a sports team to be able to automatically analyse when and where the certain activities are performed by the team.

Our goal for this work is to recognise five common sports types observed in an indoor arena; badminton, basketball, indoor soccer, handball, and volleyball. To overcome any privacy issues we apply thermal imaging, which produces images where pixel values represent the observed temperature. Thereby it is possible to detect people without identification. Figure 6.1 is an example of a thermal image from a sports arena.

Our hypothesis is that it is possible to classify five different sports types using a global approach based on position data only.

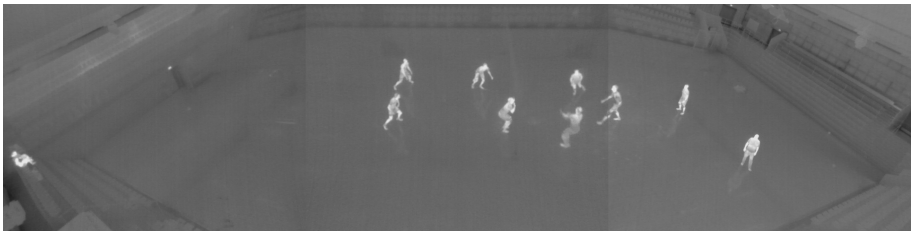


Fig. 6.1: Example of an input image.

6.2 Related work

Computer aided sports analysis has become a very popular tool, particularly in team sports, since some of the first systems were developed in the 1980's [3]. Today the computer vision community works on a large variety of automatic solutions to replace manual tracking systems. Many works have already focused on automatic tracking of sports players [4]. These systems are typically used for game statistics and evaluation of individual performances. For game analysis purposes, automatic detection of play types or special team activities in a game is also an area of interest. With applications mainly in football, offensive play strategies are tried classified based on player trajectories [5–7]. However, in situations with a high degree of interaction between people, like most types of sport, tracking is still an unsolved problem. Many classification systems therefore rely on manually annotated trajectories, in order to avoid noisy data. The work of Bialkowski et al. [8] on recognising team activities in field hockey games is related to this work in the way that they use only position data. Their approach is robust to noisy data and they test two different representations; team occupancy maps using different quantisation, and team centroids.

No previous work on recognising sports types has been based on thermal imaging. All existing works use visual cameras and a few include audio as well. For features, some works use edges that represent court lines and players. The sports categories can then be classified by edge directions, intensity, or ratio [9, 10]. Also based on the visual appearance of the court is a method that uses the dominating colours of the image as features [11]. The dominant colour can also be combined with motion features [12–14] or combined with dominant grey level, cut rate and motion rate [15]. From the visual image SURF features [16] and autocorrelograms [17] can be extracted and used for classification. A combination of colour, edges, shape and texture has also been proposed by using six of the MPEG-7 descriptors [18]. One method is based only on location data and classifies sports categories by short trajectories [19].

After feature extraction the classification methods are based on well-known methods, such as k-means and Expectation Maximization for clustering, and decision trees, SVM, Hidden Markov models, Neural Network and Naive Bayesian for classification.

As described here, most existing works are based on colour imaging, and many of them rely on the dominant colour of the fields as well as detection of court lines. These methods presume that each sports type is performed on a court designed mainly for one specific sport. In our work we aim to distinguish different sports types performed in the same arena, meaning that any information about the environment is not useful. Furthermore, due to privacy issues, we have chosen to use thermal imaging, which provides heat information only. Figure 6.1 shows an example of the thermal image, which is combined from three cameras in order to cover the full court.

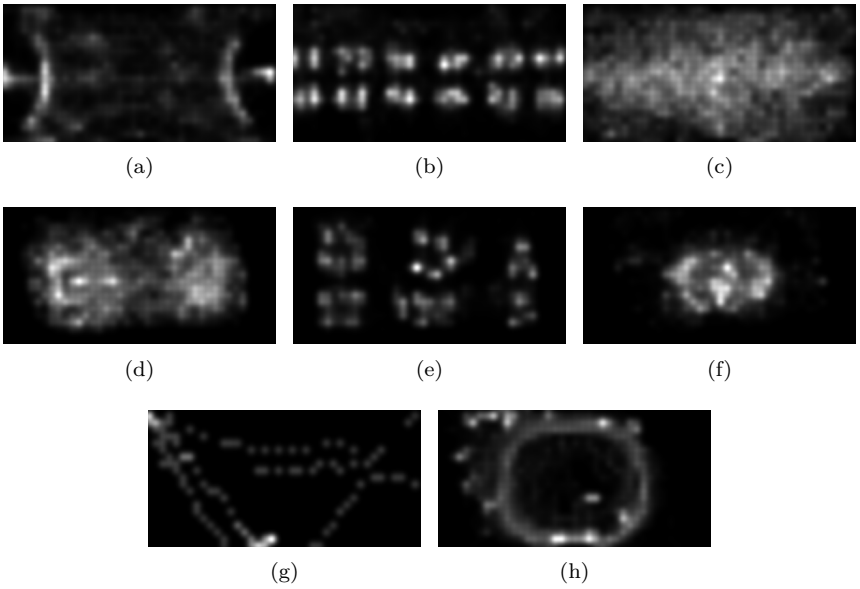


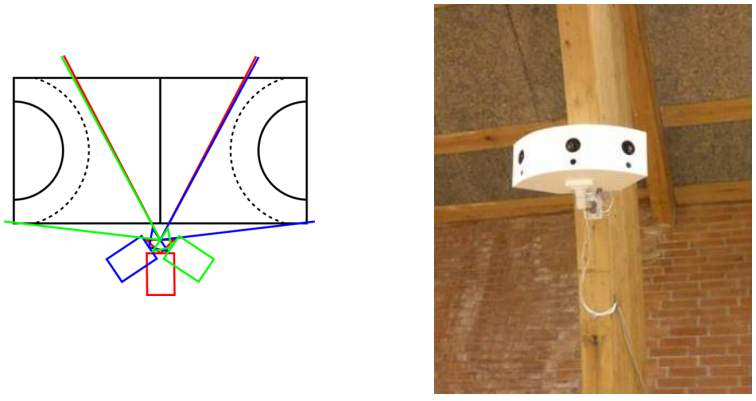
Fig. 6.2: Signature heatmaps of (a) handball, (b) badminton, (c) soccer, (d) basketball, (e) volleyball (three courts), (f) volleyball (one court), (g) miscellaneous, (h) miscellaneous.

This work is based on occupancy heatmaps, which are summations of the registered positions of people over a given time span. It is believed that a heatmap for one sports type is unique among a limited number of sports types. This approach is also robust to noisy data, due to the summations over time. Figure 6.2 shows examples of signature heatmaps, which are typical heatmaps for each sports type. Two heatmaps of miscellaneous activities are also shown. Each heatmap covers a 10-minute period.

The approach of this work is to detect individual people and use a homography to calculate their position at the court. A summation of the positions over time results in the occupancy heatmaps. These heatmaps will be classified after reducing the number of dimensions with PCA and Fisher's Linear Discriminant.

6.3 Image acquisition

In order to detect individual people without occlusions, the ideal situation would be to capture images of the scene directly from above. However, installing a camera in the ceiling above the court is generally very cumbersome and expensive and therefore not realistic for temporary installations. Therefore, it must be installed on one of the walls or rafters around the court. A standard arena used in our experiments has a court size of 40×20 metres, corresponding to a handball field, indoor soccer field, etc. As the lenses of commercial thermal cameras today have a maximum field-of-view of around 60°, more than one camera must be used to cover the entire court. The camera set-up used in this work consists of three thermal cameras placed at the same location in order to limit the work load of the set-up. The cameras are adjusted to have adjacent fields-of-view. This is illustrated in figure 6.3(a). Figure 6.3(b) shows a picture of the camera box mounted in an arena.



(a) (b)

Fig. 6.3: (a) Illustration of the camera setup. (b) Image of the camera box mounted in an arena.

Each camera is of the type Axis Q1922, which uses an uncooled microbolometer for detection. The resolution is 640×480 pixels per camera and the horizontal field-of-view is 57° per camera. This type of thermal camera is a passive sensor which captures long-wavelength infrared radiation (8-15 μm), radiated by all warm objects. The resulting image depicts the temperature of the scene [20].

To make the software system invariant to the cameras' set-up, the images are stitched together before processing. This requires the cameras to be perfectly aligned and undistorted in order to secure smooth "crossings" between two cameras. For calibration of the thermal cameras a special calibration board is

made, consisting of 5x4 small incandescent light bulbs. A thermal image of the calibration board is shown in figure 6.4. With this board it is possible to adapt

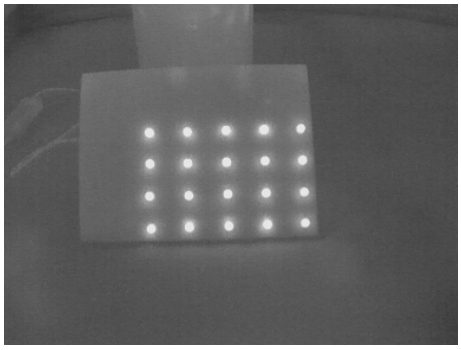


Fig. 6.4: Calibration using a board with incandescent light bulbs.

the traditional methods for estimating the intrinsic parameters of the cameras. The cameras are manually aligned horizontally so that their pitch and roll are the same. The result mimics the well-known panorama effect, but with three cameras capturing images simultaneously. An example of the resulting image is shown in figure 6.1.

6.4 Detection

For temperature controlled indoor environments, such as sports arenas, it is assumed that people are warmer than the surroundings. For thermal images this implies that people can be segmented using only thresholding of the image. Since the camera has automatic gain adjustment, the level of pixel values can change, and a constant threshold value is therefore not suitable. Instead we use an automatic threshold method. This method calculates the threshold value that maximises the sum of the entropy [21]. After binarising the image, ideally people are now white and everything else in the image is black. There are, however, challenges to this assumption. Cold/wet clothes can result in parts of people being segmented as background, meaning that one person is represented by a number of unconnected blobs. Likewise, partial occlusions will challenge the detection of individual people, as more than one person is represented in one blob. Figure 6.5 shows some examples of the challenges after binarisation of the thermal images.

The next two sections will present methods for reducing these problems.

6.4.1 Occlusion handling

Two cases of occlusions are handled; People standing behind each other seen from the camera and people standing close to each other horizontally, e.g., in

Fig. 6.5: Examples of thermal sub images and the resulting binary images.

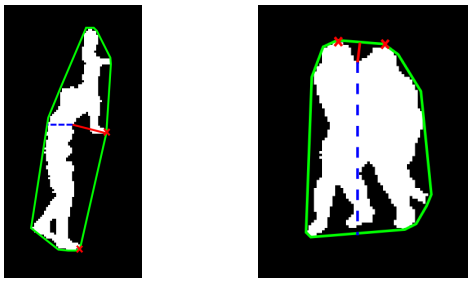
a group. In the first case, the detected blobs will appear taller than an object corresponding to one person. Maximum height of a person at each position will be determined during the initialisation. If the blob exceeds the maximum height, it should be split horizontally. Since one or more people are partly occluded in these cases, it is not trivial to find the right place to split the blob. However, we often observe some narrowing or gap in the blob contour near the head of the person in front, which is generally the best point to split from. This point is found by analysing the convex hull and finding the convexity defects of the blob. Of all the defect points, the point with the largest depth and a given maximum absolute gradient should be selected, meaning that only defects coming from the side will be considered, discarding e.g. a point between the legs. An example is shown in figure 6.6(a), where the convex hull is drawn with green and the largest convexity defect is shown with red. From the defect point the blob is split horizontally, shown with the blue dashed line.

In the second case, wide blobs containing two or more persons standing next to each other must be identified. The height/width ratio and the perimeter are here considered as done in [2]. If the criteria are satisfied, the algorithm should try to split the blob. For this type of occlusion, it is often possible to see the head of each person, and split the blob based on the head positions. Since the head is narrower than the body, people can be separated by splitting vertically from the minimum points of the upper edge of a blob. These points can be found by analysing the convex hull and finding the convexity defects of the blob as shown in figure 6.6(b). The convex hull is drawn with green, and the convexity defect of interest is drawn with red. The convexity defect start and end points, shown with red crosses, must both be above the defect point in the image, therefore other defects are discarded. The split line is shown with the blue dashed line.

More examples of blobs to be split are shown in figure 6.7(a) and 6.7(b).

6.4.2 Joining of blobs

Another challenge to the detection of individual people is separation of one person into several blobs. This happens when part of the body has a lower temperature, e.g., caused by loose, wet clothes or insulating layers of clothes. In order to detect the correct position of a person, the bottom part of the body must be identified. We adapt here the method presented in [1]. Each detected blob is considered a person candidate, and the probability of being a true person is tested. From the bottom position of the blob, a bounding box of



(a) (b)

Fig. 6.6: Example of how to find the split location of tall or wide blobs.

the expected height of an average person and the width being one third of the height is generated. The probability for the candidate being true depends on the ratio of white pixels (r) in the rectangle. By tests it is found that most true detections have a ratio between 30 % and 50 %, while less than 1 % of the true detections lie below 20 % or above 70 %. We choose to discard all detections below 20 % and assign a value between 0.8 and 1 to all other detections:

$$w_p(i) = \begin{cases} 0, & \text{if } r < 20\% \\ 0.8, & \text{if } r > 60\% \\ 0.9, & \text{if } r < 30\% \parallel 50 < r < 60\% \\ 1, & \text{otherwise} \end{cases} \tag{6.1}$$

This approach will reduce the detections of small body parts, as well as many non-human objects. But a lot of false candidates will still exist. Many of them contain part of a person and overlap in the image with a true candidate. Due to the possibility of several candidates belonging to the same person, the overlapping rectangles must be considered. By tests from different locations and different camera placements, it is found that if two rectangles overlap by more than 60 %, they probably originate from the same person, or from reflections of that person. As only one position should be accepted per person, only one of the overlapping rectangles should be chosen. Due to low resolution images compared to the scene depth, cluttered scenes, and no restrictions on the posture of a person, the feet of a person can not be recognised from the blobs. Furthermore, due to the possibility of reflections below a person in the image, it can not be assumed that the feet are the lowest point of the overlapping candidates. Instead, the best candidate will be selected on the highest ratio of white pixels, as the probability of false candidates is lower here. The probabilities assigned to the approved candidates will be used later, when registering the positions of people.

Figure 6.7(c) shows three examples of blobs being joined to one detected person.

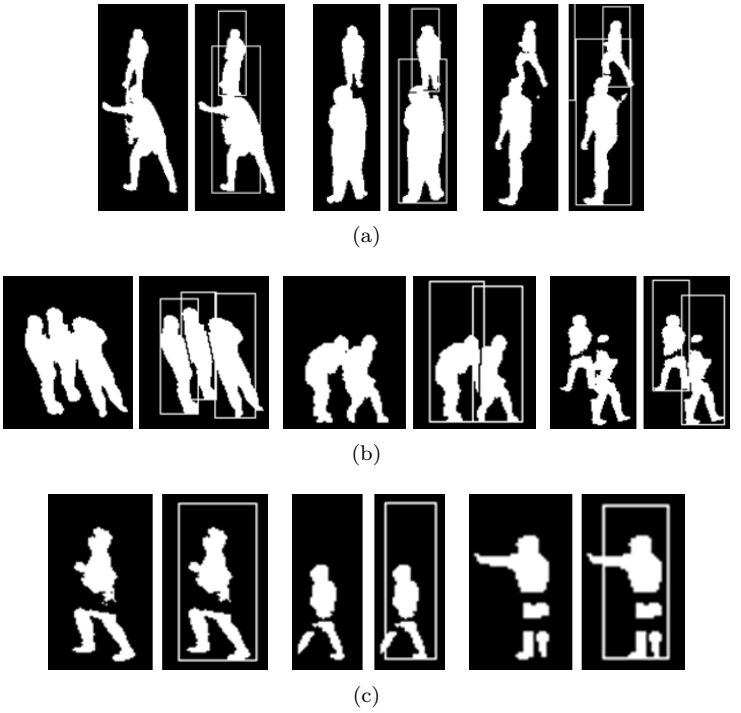


Fig. 6.7: Illustration of how blobs can be split or joined to single persons.

6.4.3 Region of interest

As spectators, coaches and other persons around the court are of no interest in this work, the image must be cropped to the border of the court before processing. Since each sports type has its own court dimensions, a single choice of border is not feasible. Handball and soccer are played on a 40×20 metres court, which is also the maximum court size in the observed arena. The volleyball court is 18×9 metres, plus a free zone around the court, which is minimum 3 metres wide, and the standard basketball court is 28×15 metres. Badminton is played on up to six adjacent courts of 13.4×6.1 metres. The court dimensions and layout in relation to each other are illustrated in figure 6.8. On the arena floor all court lines are drawn on top of each other, but here we have split it into two drawings for better visibility. Note that volleyball can be played on either three courts without free zones or on one court including the free zone.

During basketball and volleyball matches coaches and substitutes will be sitting within the dimensions of the handball court, and would be unwanted detections if we cropped only to the largest court dimensions. Considering the illustrated court dimensions it is therefore decided to operate with two different court sizes, 40×20 metres and 28×15 metres. In test cases it is of course not

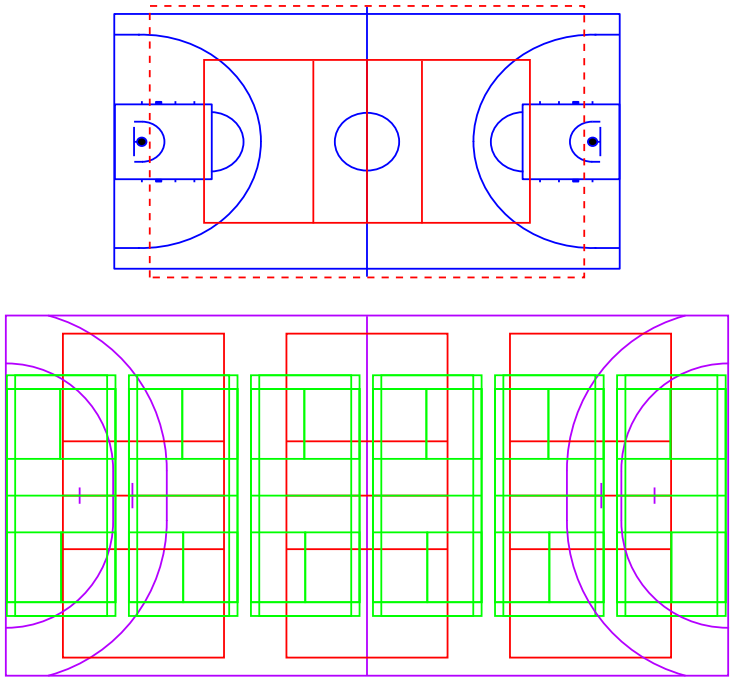


Fig. 6.8: The outlines of the different courts illustrated. Red: volleyball, blue: basketball, purple: handball (and soccer), green: badminton. Drawn in two figures to increase the visibility.

known which sport is performed and thereby not known which court size to choose. Instead both options will be tried out for all data. The classification process will be further described in section 6.5.

6.4.4 Occupancy heat maps

The position of each person is registered in image coordinates as the bottom centre of the bounding box. The position then needs to be converted to world coordinates using a homography. Since the input image is combined from three cameras, each observing the left, middle or right part of the court, at least one homography for each camera is needed to calculate the transformation. This assumes that the cameras are perfectly rectified. For better precision, we instead divide the court into 5×5 metres squares for each of which we calculate a homography. The corresponding points in image and world coordinates for each five metres in both x- and y-direction are found during an initialisation. This initialisation must be performed one time for each set-up. In addition to finding the mapping between image and world coordinates, we also find the correlation between peoples' real height and their height in the images, corresponding to their distance to the camera. Furthermore, as the cameras are fixed relative to

each other in one box and then tilted downwards when mounted in arenas, the result is that people in the image are more tilted the further they get from the image centre along the image x-axis. This means that a person's pixel height can not always be measured vertically in the image. Therefore, the calibration must include the angle of a person standing upright at predefined positions on the court.

Figure 6.9 illustrates the result of an initialisation procedure. For each position in the 5×5 metres grids a person stands upright, while the image coordinates of the feet and head are registered, resulting in a white line in the image, which gives the angle and height of a person. The world coordinate of the feet are also registered as well as the person's real height in metres.

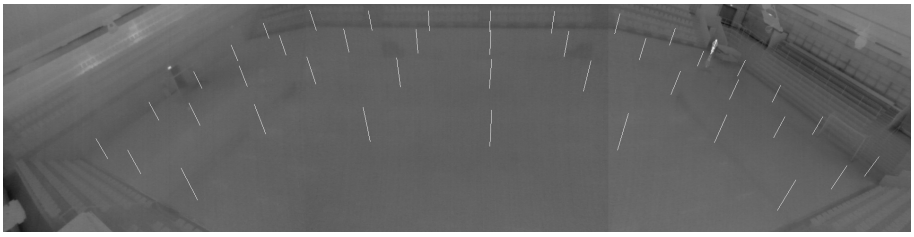


Fig. 6.9: Illustration of the initialisation process, each white line represents a standard person at that position.

The four corner points of each square are used to calculate the homographies, making it possible to map all image coordinates to world coordinates. Using interpolation from the corner points, an angle and maximum height are calculated for each pixel.

When mapping the observations to a world-coordinate view of the court we need to represent the physical area of a person. A standard person is chosen to be represented by a 3-dimensional Gaussian distribution with a standard height of 1, corresponding to 1 person, and a radius corresponding to 1 metre for 95% of the volume. To take into account the uncertainty of the detections, the height of Gaussian distributions will be scaled by the probability factor w_p, described in section 6.4.2.

Figure 6.10 shows an example of the occupancy calculated for a single frame. Six people are detected with different certainty factors.

The final occupancy heatmaps, as shown in figure 6.2, are constructed by adding up the Gaussians over time. The time span for each heatmap should be long enough to cover a representative section of the games and still short enough to avoid different activities to be mixed together. To decide on the time span, a study has been conducted between 5-, 10-, 20- and 30-minutes periods. An example of four heatmaps with the same end-time is shown in figure 6.11.

The comparison in figure 6.11 illustrates the situation where a handball team starts with exercises and warm-up, before playing a short handball match. The end-time for each heatmap is the same. The 30-minute period (figure

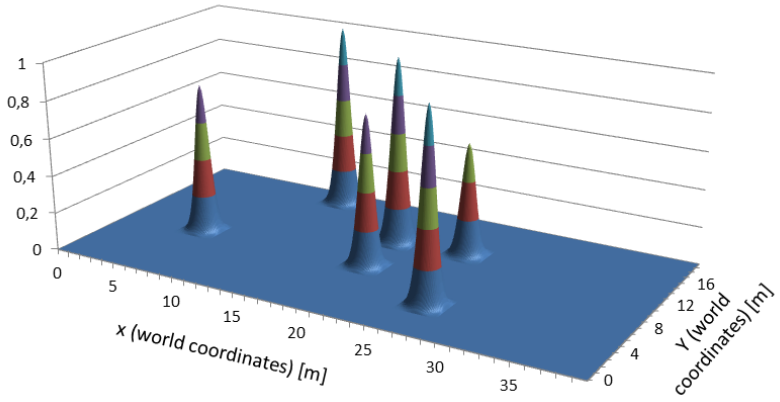


Fig. 6.10: Occupancy for one single frame. Each person is represented as a Gaussian distribution.

6.11(d)) is too long, the warm-up and game is mixed together such that no activity is recognisable. Between the 5-, 10- and 20-minute periods the 10-minute period (figure 6.11(b)) shows the most clear pattern. The same is observed in comparisons for other sports activities, therefore it is chosen to let each heatmap cover 10 minutes. We will shift the starting time 5 minutes each time, so that the periods overlap and the resolution of classifications will be 5 minutes.

6.5 Classification

Each heatmap is an images with a resolution of 200×400 pixels, thus it can be considered a sample in an 80,000-dimensional space. Principal Component Analysis (PCA) is a well-known method for dimension reduction, but since it uses non-labeled data and seeks the dimensions with largest variance between all samples, there is a risk that the differences between classes are not conserved. Fischer's Linear Discriminant (FLD) seeks the directions that are efficient for discrimination between classes [22]. However, using FLD introduces the small sample size problem: In order to have a non-singular within-class scatter matrix (S_W) it is necessary to have more samples than dimensions. As we have an 80,000-dimensional space, it is not realistic to have a sample size of $n > 80{,}000$. In order to solve this problem we will adapt the idea of Fisherfaces for face recognition [23]: First, project the sample vectors onto the PCA space of r dimensions, with $r \leq rank(S_W)$ and then compute the Fisherimage in this PCA space.

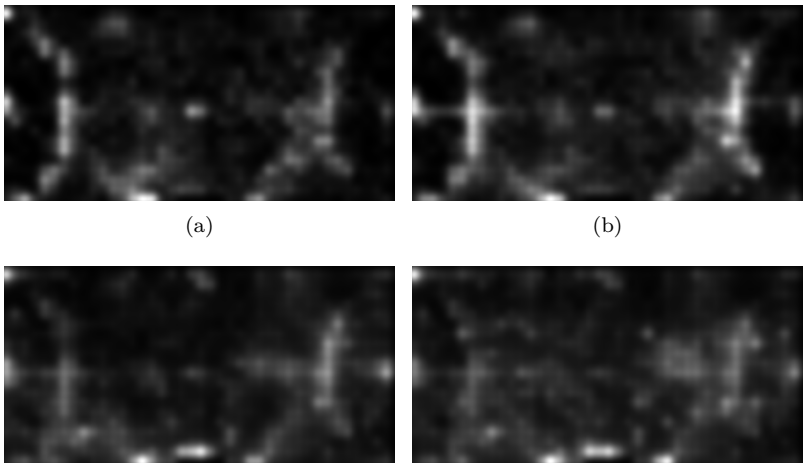


Fig. 6.11: Heatmaps with same end-time and different time span: (a) 5 minutes, (b) 10 minutes, (c) 20 minutes, (d) 30 minutes.

6.5.1 Dimensionality reduction

Dimensionality reduction based on PCA is performed by pooling all training samples and calculating the directions with largest variance. The PCA will only have as many non-zero eigenvalues as the number of samples minus one, which will be significantly less than the original 80,000 dimensions. We choose to reduce the space to the 20 dimensions with largest eigenvalues. Even though this is a reduction by 4000 times in the number of dimensions 73 % of the total variance is still preserved. All heatmaps are projected to the new 20-dimensional space before further processing.

6.5.2 Fischer's Linear Discriminant

The optimal projection of the data is found using Fisher's Linear Discriminant, such that the ratio of the between-class scatter S_B and the within-class scatter S_W is maximised:

$$W_{opt} = arg \max_W \frac{|W^T S_B W|}{|W^T S_W W|} \tag{6.2}$$

where W_{opt} is a matrix with orthonormal columns, consisting of the set of generalised eigenvectors of S_B and S_W corresponding to the m largest eigenvalues. There are at most $c-1$ non-zero generalised eigenvalues, where c is the number of classes.
The between-class scatter matrix S_B and the within-class scatter matrix S_W

are defined as:

$$S_B = \sum_{i=1}^{c} N_i(\mu_i - \mu)(\mu_i - \mu)^T \tag{6.3}$$

$$S_W = \sum_{i=1}^{c} \sum_{x_k \in X_I} N_i(\mu_i - \mu)(\mu_i - \mu)^T \tag{6.4}$$

where μ_i is the mean image of class X_i, and N_i is the number of samples in class X_i [23].

6.5.3 Final classification

In the training phase, all data is projected to the new space found by FLD, and the mean coordinate for each class is calculated. When testing a new sample, the sample to classify is projected to the same space, and the nearest class is found using the Euclidean distance.

We use video from daily activities in a public sports arena, which includes a lot of undefined activities. Besides the five traditional sports types we therefore define a category of miscellaneous activities. This can include everything from specific exercises and warm-up, to cleaning the floor and an empty arena. This category will be trained as a class in the same way as each sports type. Since miscellaneous contains very different heatmaps, it could be argued that this class will end up as a mean image of all other classes. However, by treating it as a class like the other sports types, the FLD will take care of projecting the data to a space that, as far as possible, discriminates the classes.

We will use two different court dimensions for tests, as described in section 6.4. The final classification for each time span should therefore choose between the classifications of these two heatmaps. If they agree on the same class, the final classification is simply that class. If one heatmap is classified as a sports type, while the other is classified as miscellaneous, the final classification will choose the regular sports type, as it is assumed that it found the correct court size. If the heatmaps are classified as different sports types, the sample with shortest distance to the class mean will decide the classification.

6.6 Experiments

19 days of data has been captured in order to test the classification approach. Capturing from 7am to 11pm this is a total of 304 hours of recordings, of which people are present in the arena in 163 hours and 141 hours are empty. Video from the first week (7 days) is used for training data and the rest (12 days) is used for test. This approach is challenging, since the variety in the play can be large between different sessions. Many undefined activities are observed during a day, from warm-up and exercises, to more passive activities,

such as transitions between teams, "team meetings", cleaning, etc. Only well-known sports types performed like in matches will be used for classification. Exercises related to a specific sport, such as practising shots at goal, will not be considered a specific sports type, but will be counted as miscellaneous activity. We do, however, allow variety in the play, such as different number of players and different number of courts in use for badminton and volleyball.

The sports types that are observed during both weeks and will be used in this work are badminton, basketball, indoor soccer, handball, and volleyball. As shown in figure 6.8, two different layouts of volleyball courts are observed, one with only one court in the middle of the arena (drawn on upper part of fig. 6.8 and denoted volleyball-1) and the other version which fit three volleyball courts playing in the opposite direction (drawn with red on lower part of fig. 6.8 and denoted volleyball-3). These will be treated as two different classes, both referring to volleyball. This results in seven classes to classify, including miscellaneous.

For training and test of each sports type we use all heatmaps that are manually labelled to be a regular performed sport. In order to have a significant representation of the regular sports types in the total dataset we discard most of the empty hours and use only a few heatmaps from empty periods. Furthermore, the rest of the miscellaneous heatmaps are chosen as samples that represent the various kinds of random activities that take place in the arena. The number of heatmaps used for each class is listed in table 6.1.

Category	**Training heatmaps**	**Test heatmaps**
Badminton	35	19
Basketball	16	76
Soccer	20	36
Handball	18	15
Volleyball-1	33	20
Volleyball-3	15	8
Misc.	163	212
Total	300	386

Table 6.1: Data set used for training and test.

In order to test the system under real conditions, which will be continuous video sequences of several hours, we do also perform a test on video captured on one day continuously from 7am to 11pm. This video contains recordings of volleyball, handball and soccer, as well as miscellaneous activities. The training data described in table 6.1 is used again for this test. At last, we test our algorithm on a publicly available dataset from a handball game, while still using our own videos for training data. This will prove the portability of our method to other arenas and set-ups.

6.6.1 Results

12 days test

Table 6.2 shows the result for the first test with data from 12 days. The ground truth is compared with the classification.

Truth \ Classified	Badm.	Bask.	Soc.	Hand.	Volley-1	Volley-3	Misc.
Badminton	**17**	0	0	0	0	0	2
Basketball	0	**69**	0	0	1	0	6
Soccer	0	0	**30**	0	4	0	2
Handball	0	0	0	**15**	0	0	0
Volleyball-1	0	0	0	0	**20**	0	0
Volleyball-3	0	0	0	0	0	**4**	4
Misc.	0	14	2	1	2	2	**191**

Table 6.2: Classification result for data samples from one week. The number of heatmaps classified in each category.

This results in an overall true positive rate of 89.64 %. This result is very satisfying, considering that we classify seven classes based only on position data.

A low number of 14 heatmaps are wrongly classified as miscellaneous instead of the correct sports type. Four of them are from videos where only one of the three volleyball courts is used, and this situation is not represented in the training data. The error could therefore be reduced by capturing more training data. Of the basketball videos, a few heatmaps represent periods with unusually many players on the court, resulting in a different activity pattern and therefore they are classified as miscellaneous. Fourteen heatmaps manually labelled as miscellaneous are automatically classified as basketball. These heatmaps are borderline situations where exercises highly related to basketball are performed, and it could therefore be discussed whether these should be labelled basketball or miscellaneous. The same happens for a few miscellaneous heatmaps, classified as other sports types due to exercises highly related to the sport. Four heatmaps representing soccer are misclassified as volleyball played on the centre court. Inspecting these images, there are some similarities between the sports, depending on how they are performed.

Full day test

The result of classifying one full day from 7am to 10pm is illustrated in figure 6.12 with each colour representing a sports type and grey representing miscellaneous activities (including empty arena). The ground truth is shown in the upper row and automatic classification is shown in the bottom row.

The result is very promising, showing that of the total of 191 heatmaps that are produced and classified for the full day, 94.24 % are correctly classified.

The green periods illustrate volleyball matches. Before these matches there is a warm-up period, where short periods of exercises are confused with basket-

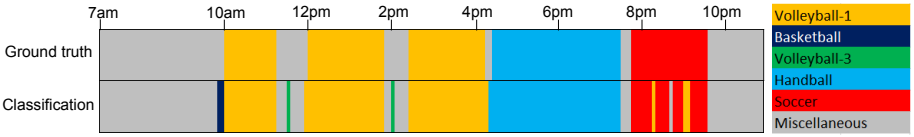


Fig. 6.12: Comparison of ground truth and classification of video from one full day.

ball or volleyball played on the three courts. The last case is obvious, because some of their warm-up exercises include practising volleyball shots in the same direction as volleyball is normally played using the three courts. This test does also show like the previous test that soccer can be misclassified as volleyball in a few situations.

The results from this test show that our approach works very satisfying even for the challenging situation of a full day's video, the true positive rate is indeed better than what was obtained in the first test.

CVBASE dataset

The last test performed is classification of the sport from a publicly available dataset. In order to do that, we need a dataset with at least 10 minutes continuously recording of one of the five sports type considered in this paper. Furthermore, calibration data must be available, so that positions can be obtained in world coordinates. One suitable dataset is found, which is the handball dataset from CVBASE 06 [24]. This includes annotation of position data in world coordinates for seven players (one team) for 10 minutes of a handball match. Since we want to test the classification algorithm specifically, we use these annotations as input data instead of modifying our detection algorithm to work on RGB video. However, as we need position data from the players of both teams, we flip all positions along the x-axis (longest dimension of the court) and add those positions in order to represent the other team. The resulting heatmap for the 10-minute period is shown in figure 6.13.

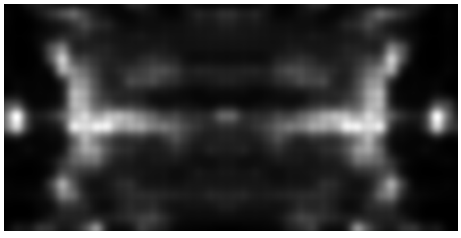


Fig. 6.13: Heatmap for the 10 minutes annotated positions of the CVBASE 06 handball dataset.

Using the one week training data from our own recordings, this 10 minute period is correctly classified as handball. This proves the portability of our

approach to other arenas and camera set-ups.

6.6.2 Comparison with related work

A comparison of our results with the reported results in related work is listed in table 6.3. It should be noted that each work has its own data set, making it hard to compare the results directly. All related works use normal visual cameras, where we use thermal cameras. In addition to that, most work use video from different courts for each sports type, where we use video from one multi-purpose indoor arena.

Reference	Sports types	Video length	Result
Gibert et. al [14]	4	220 min.	93 %
Mohan and Yegn. [9]	5	5 h. 30 min.	94.4 %
Lee and Hoff [19]	2	Approx. 1 hour	94.2 %
Li et. al [16]	14	114 hours	88.8 %
Mutch. and Sang. [11]	20	200 min.	96.65 %
Sigari et. al [15]	7	(104 video clips)	78.8 %
Wang et. al [13]	4	(173 test clips)	88 %
Wang et. al [12]	3	16 hours	100 %
Watcha. et. al [17]	7	233 min.	91.1 %
Xu et. al [18]	4	1200 frames	N/A
Yuan and Wan [10]	5	N/A	97.1 %
Our work	5	65 hours	89.64 %

Table 6.3: Data set used for training and test.

Our result is comparable with the related work using an equal number of sports types. It is also seen that we test on a large amount of data compared to other works.

6.7 Conclusion

The work presented here shows that it is possible to classify five different sports types based only on the position data of people detected in thermal images. Heatmaps are produced by summarising the position data over 10-minute periods. These heatmaps are projected to a low-dimensional space using PCA and Fischer's Linear Discriminant. Our result is an overall recognition rate for five sports types of 89.64 %. This is a very promising result, considering that our work is the first to use thermal imaging for sports classification. Furthermore, we use video from the same indoor arena, meaning that no information about the arena can be used in the classification. Our detection method is rather simple and the registered positions of people can be noisy, but since the classification method relies on summarised positions over time, the approach is robust to the noisy data.

For this work we have concentrated on sport played in match-like situations. Problems could rise if trying to classify a video of sport played in the opposite direction of usual, e.g. on half the court, or if trying to classify exercises related to one sports type. To overcome these limitations future work will investigate the possibility of including local features. These could be clues from short trajectories, such as speed and path length and straightness to overcome these limitations. In relation to this, it could also be possible to extend the work to classify play types or other shorter activities within a sports game.

References

[1] R. Gade, A. Jørgensen, and T. B. Moeslund, "Long-term occupancy analysis using graph-based optimisation in thermal imagery," in *CVPR*, 2013.

[2] ——, "Occupancy analysis of sports arenas using thermal imaging," in *Proceedings of the International Conference on Computer Vision and Applications*, feb. 2012.

[3] H. Steiner and M. Butzke, "Casa — computer aided sports analysis," in *Hector*, B. Krause and A. Schreiner, Eds. Springer Berlin Heidelberg, 1988, pp. 182–185.

[4] S. Barris and C. Button, "A review of vision-based motion analysis in sport," *Sports Medicine*, vol. 38, no. 12, pp. 1025–1043, 2008.

[5] J. Varadarajan, I. Atmosukarto, S. Ahuja, B. Ghanem, and N. Ahujad, "A topic model approach to representing and classifying football plays," in *British Machine Vision Conference*, 2013.

[6] R. Li, R. Chellappa, and S. Zhou, "Learning multi-modal densities on discriminative temporal interaction manifold for group activity recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[7] R. Li and R. Chellappa, "Recognizing offensive strategies from football videos," in *IEEE International Conference on Image Processing*, 2010.

[8] A. Bialkowski, P. Lucey, P. Carr, S. Denman, I. Matthews, and S. Sridharan, "Recognising team activities from noisy data," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013.

[9] C. Krishna Mohan and B. Yegnanarayana, "Classification of sport videos using edge-based features and autoassociative neural network models," *Signal, Image and Video Processing*, vol. 4, pp. 61–73, 2010.

[10] Y. Yuan and C. Wan, "The application of edge feature in automatic sports genre classification," in *IEEE Conference on Cybernetics and Intelligent Systems*, 2004.

[11] P. Mutchima and P. Sanguansat, "TF-RNF: A novel term weighting scheme for sports video classification," in *IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC)*, 2012.

[12] J. Wang, C. Xu, and E. Chng, "Automatic sports video genre classification using Pseudo-2D-HMM," in *18th International Conference on Pattern Recognition (ICPR)*, 2006.

[13] D.-H. Wang, Q. Tian, S. Gao, and W.-K. Sung, "News sports video shot classification with sports play field and motion features," in *International Conference on Image Processing (ICIP)*, 2004.

[14] X. Gibert, H. Li, and D. Doermann, "Sports video classification using HMMS," in *International Conference on Multimedia and Expo (ICME)*, 2003.

[15] M. Sigari, S. Sureshjani, and H. Soltanian-Zadeh, "Sport video classification using an ensemble classifier," in *7th Iranian Machine Vision and Image Processing (MVIP)*, 2011.

[16] L. Li, N. Zhang, L.-Y. Duan, Q. Huang, J. Du, and L. Guan, "Automatic sports genre categorization and view-type classification over large-scale dataset," in *17th ACM international conference on Multimedia (MM)*, 2009.

[17] N. Watcharapinchai, S. Aramvith, S. Siddhichai, and S. Marukatat, "A discriminant approach to sports video classification," in *International Symposium on Communications and Information Technologies (ISCIT)*, 2007.

[18] M. Xu, M. Park, S. Luo, and J. Jin, "Comparison analysis on supervised learning based solutions for sports video categorization," in *IEEE 10th Workshop on Multimedia Signal Processing*, 2008.

[19] J. Y. Lee and W. Hoff, "Activity identification utilizing data mining techniques," in *IEEE Workshop on Motion and Video Computing (WMVC)*, 2007.

[20] R. Gade and T. B. Moeslund, "Thermal cameras and applications: a survey," *Machine Vision and Applications*, 2013.

[21] J. Kapur, P. Sahoo, and A. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Computer Vision, Graphics, and Image Processing*, vol. 29, no. 3, pp. 273 – 285, 1985.

[22] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, 2001.

[23] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *PAMI*, vol. 19, no. 7, pp. 711 –720, jul 1997.

[24] M. B. Janez Pers and G. Vuckovic. (2006) CVBASE 06 Dataset. [Online]. Available: http://vision.fe.uni-lj.si/cvbase06/dataset.html

Chapter 7

Classification of Sports Types from Tracklets

Rikke Gade and Thomas B. Moeslund

The paper is presented at
KDD workshop on Large-scale Sports Analytics, August 2014.

© 2014
The layout has been revised.

Abstract

Automatic analysis of video is important in order to process and exploit large amounts of data, e.g. for sports analysis. Classification of sports types is one of the first steps towards a fully automatic analysis of the activities performed at sports arenas. In this work we test the idea that sports types can be classified from features extracted from short trajectories of the players. From tracklets created by a Kalman filter tracker we extract four robust features; Total distance, lifespan, distance span and mean speed. For classification we use a quadratic discriminant analysis. In our experiments we use 30 2-minutes thermal video sequences from each of five different sports types. By applying a 10-fold cross validation we obtain a correct classification rate of 94.5 %.

7.1 Introduction

Manual analysis of video is very time consuming and expensive. Automating the analysis will enable a significantly higher amount of data to be processed and exploited for systematic analysis of, e.g., sports activities. The interest in sports analytics has grown significantly recently as governments, broadcasters, coaches, etc. see great potential in the data. In this work we focus on automatic recognition of sports types. For large amounts of video, this step will help separating the data into sequences of well-known sports types. Furthermore, for multi-purpose indoor arenas as well as outdoor fields, it can be of great interest to get a better knowledge of the use of the facilities, without having to perform manual annotation. We have previously proposed a method for activity recognition based on heatmaps produced from summed position data [1]. In this work we will try to estimate which type of sport is being performed based on motion features extracted from tracklets.

Previous work on sports type recognition has often been based on the visual appearance of the court, such as court lines and dominant colour of the field [2–4]. The dominant colour has also been combined with motion features, such as camera/background motion [5, 6] or direction of motion vectors in image blocks [7]. In this work we will classify different sports types performed in the same indoor multi-purpose arena. The appearance of the court will therefore not be useful for classification. Furthermore, we use a static camera setup with thermal cameras, eliminating both camera motion features and any colour features. Thermal cameras are chosen in order to minimise the privacy issues of capturing video in public sports arenas.

Most relevant to this work then is mainly two papers. Lee and Hoff [8] detect players and use trajectory segments of three seconds from which they extract and test eight features based on speed, direction and path length. They find that two features maximises the classification accuracy. These features are average speed and the ratio of the overall distance to the path length. Using k-means clustering and decision tree classification, they achieve 94.2% accuracy.

However, they test on only two sports types; Ultimate Frisbee and volleyball. Whether these two features will be sufficient to discriminate a larger set of sports types is therefore unknown. Gade and Moeslund [1] proposed sports type recognition based on classification of heatmaps produced from position data. The heatmaps are projected to a low-dimensional discriminative space using Fischers Linear Discriminant and new instances are classified as the nearest cluster. In this work five different sports types are classified with a precision of 90.8 %. Limitations of this work include the dependency on scale, direction and location on the field. To overcome these limitations, we will in this work extract local features, which are invariant to the position and direction of play. Based on trajectories (tracklets) from each player, motion features are extracted and used for classification.

In the remaining part of this paper, section 7.2 will describe the tracking algorithm used to produce tracklets, after which we choose the features to extract in section 7.3. In section 7.4 the classification approach is described, before the experiments and results are presented in section 7.5, and finally the conclusion is found is section 7.6.

7.2 Tracking

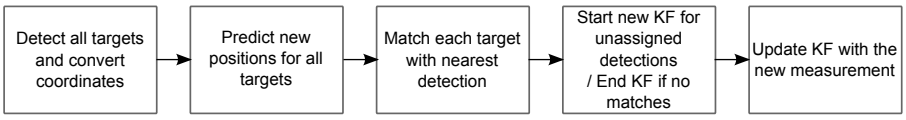


Fig. 7.1: Illustration of the tracking framework which is run for every frame.

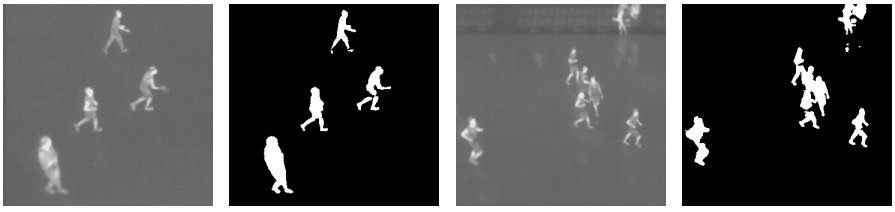
To analyse the motion of people, we need their trajectories. Tracking multiple people through interactions, occlusion and complex motion is a problem no existing methods solve automatically yet. Instead, we here aim to obtain short but reliable trajectories (tracklets), from which we can estimate motion features.

In thermal images the appearance information of people is very sparse, as only temperature is measured. When people are observed from a distance of several metres, small differences in temperature patterns will not be visible, hence people will appear as grey blobs of similar temperature. A cropped input image is shown in figure 5.5. The similar appearance of people must be considered when choosing the tracking scheme.

We choose a classic approach based on the Kalman filter [9]. This method is one of the predict-match-update schemes, which predicts the next position of the object from the previous state (described by, e.g., position and velocity), then updates the estimate when a (probably noisy) measurement is obtained. Using Kalman filtering for multi-target tracking can be done by assigning a new Kalman filter for each new target, however, it implies some reasoning

for assigning each detection to the right tracker. This is here determined by the shortest Euclidean distance within a given threshold. If a detection is not assigned to a tracker, a new Kalman filter is started. Likewise, if no detections are assigned to a tracker in n consecutive frames, the Kalman filter is terminated. n is experimentally set to 10 frames. Figure 7.1 illustrates the tracking process performed for each frame.

The first step illustrated in figure 7.1 is the detection of all targets in the frame. We use thermal imaging in an indoor arena, making it reasonable to assume that people appear warmer than a static background. The main step in our detection algorithm is therefore an automatic thresholding of the image. Figure 7.2 illustrates this step.



(a) (b) (c) (d)

Fig. 7.2: Example of a thermal input images (a) and (c), and the corresponding binary image after automatic thresholding (b) and (d).

In figure 7.2(a) and 7.2(b) people are nicely separated and easily segmented. However, occlusions can cause problems for detecting individual people, as shown in figure 7.2(c) and 7.2(d). To overcome some of these problems we try to detect these blobs containing more than one person and split them either horizontally or vertically. Further details on these procedures can be found in [10].

To be independent of the image perspective we transform the detected positions of people in the image into world coordinates before tracking. This is done by applying a homography matrix, calculated during initialisation.

Terminating the tracks with no possibility of re-identification later will naturally lead to more split trajectories. But as the identity of players has no role in this work, it is preferable to have short reliable tracklets instead of trying to resolve complex situations with a higher probability of false tracks.

Figure 7.3 shows examples of typical trajectories extracted from 2-minutes video sequences of each sports type. Each tracklet is assigned a random colour and are presented in world coordinates (top view of the court).

7.3 Features

From the trajectories we will extract features representing the typical type of motion for each sports type. We consider the following five types of features:

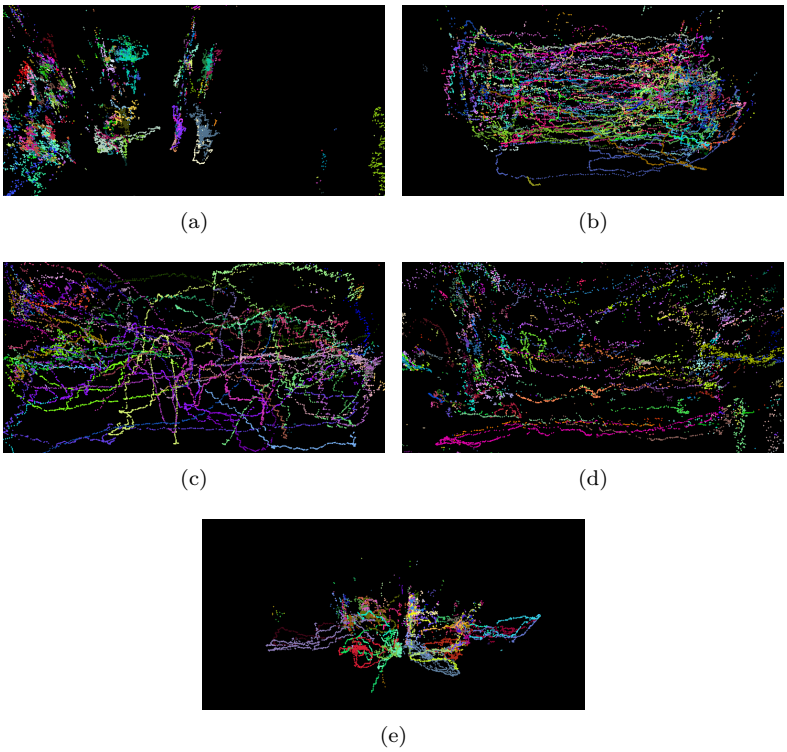


Fig. 7.3: Tracklets from a 2-minute period of (a) badminton, (b) basketball, (c) soccer, (d) handball, and (e) volleyball.

- **Speed:** Mean speed, acceleration, jerk
- **Direction:** Distribution of directions, change in direction
- **Distance:** Euclidean distance from start to end point, total distance travelled, largest distance span between two points
- **Motion:** Total motion per frame
- **Position:** Distance between people, area covered

As discussed in the introduction, we aim to find a few simple features, which should be invariant to the size and direction of the court, the position of the players according to the court and to the direction of play. The features must be robust to noisy detections and tracking errors as well. Acceleration, jerk, change in direction and euclidean distance from start to end point are all discarded because they are easily affected by tracking noise. The distribution of direction depends on the direction/rotation of play, and is therefore discarded.

The motion and position features are discarded as they depend on number of people present on the court, as well as size of the play area. Hence, we end up with the following four features calculated for each tracklet:

Lifespan [frames] is measured in number of frames before the tracklet is terminated. This feature implicitly represents the complexity of the sequence; the lifespan of each tracklet will be shorter when the scene is highly occluded:

$$ls = n_{end} - n_{start} \tag{7.1}$$

where n is the frame number.

Total distance [m] represents the total distance travelled, measured as the sum of frame-to-frame distances in world coordinates:

$$td = \sum_{i=0}^{ls-1} d(i, i+1) \tag{7.2}$$

where d is the Euclidean distance function.

Distance span [m] is measured as the maximum distance between any two points of the trajectory. This feature is a measure of how far the player move around at the court:

$$ds = \max(d(i,j)), \quad 0 < i < ls,\ 0 < j < ls \tag{7.3}$$

Mean speed [m/s] is measured as a mean value of the speed between each observation:

$$ms = \frac{td \cdot n_{seq}}{ls \cdot t} \tag{7.4}$$

where t is the duration of the video sequence in seconds, and n_{seq} the duration of the sequence in number of frames.

For each video sequence used in the classification, we will use the mean value for each feature and combine the features with equal weighting. We test all combinations of the features described above, from using a single feature to using all four. We find that the best results are obtained when using all four features, indicating that none of them are redundant or misleading.

7.4 Classification

For the classification task we choose to use a supervised learning method. We provide labelled training data and aim to find a function that best discriminates

the different classes. For this purpose we apply discriminant analysis with both a linear (LDA) and a quadratic discriminant function (QDA). The simpler linear function LDA estimates the planes in the n-dimensional space that best discriminates the data classes [11]:

$$g_l(\mathbf{x}) = w_0 + \sum_{i=1}^{n} w_i x_i \tag{7.5}$$

where the coefficients w_i are the components of the weight vector $\mathbf{w}$ and n is the number of dimensions of the space. The quadratic function estimates an hyperquadric surface:

$$g_q(\mathbf{x}) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} x_i x_j \tag{7.6}$$

The best choice of discriminant function depends on the nature of the data, and we will test both linear and quadratic functions.

Each of the five sports types is considered a class. In the classification phase, each new sample is assigned to the class with smallest misclassification cost.

In this work we do not consider undefined activities, such as warm up and exercises, as the number and variety of these activities might be unlimited, thus not representable in a single class.

7.5 Experiments

Truth \ Classified as	**Badminton**	**Basketball**	**Soccer**	**Handball**	**Volleyball**
Badminton	**29**	0	0	1	0
Basketball	0	**27**	2	0	1
Soccer	0	0	**29**	0	1
Handball	0	0	1	**27**	0
Volleyball	0	2	0	0	**26**

Table 7.1: Classification results for 146 video sequences used for tests in a 10-fold cross validation.

For the experiments we use sports types which can be easily defined and thereby unambiguously annotated. From recordings made in two similar indoor multi-purpose arenas we have five well-defined sports types available: Badminton, basketball, handball, soccer, and volleyball. We use 60 minutes of video recordings from each of the five sports types and divide them into 2-minutes sequences to get a total of 150 video sequences. The experiments are run as 10-fold cross validation; using one 10th of the data for test and the remaining part for training, then repeating the process 10 times with a new data subset for test each time.

For classification we test both linear and quadratic discriminant functions as described in section 7.4. The quadratic function fits the data best and obtain a correct classification rate of 94.5 %, while the linear discriminant function has a correct classification rate of 90.4 %. Table 7.1 shows the classification result of the 146 video sequences used for tests during ten iterations, using the quadratic discriminant function.

Of the 146 sequences, 138 are correctly classified and only 8 sequences are wrongly classified, giving a total correct classification rate of 94,5 %. The errors are distributed with 1-3 wrongly classified sequences for each sports type.

Comparable work from [1] obtained a correct classification rate of 90.8 % using the same five sports types, plus a miscellaneous class.

The Kalman tracking algorithm, including detection of people, is implemented in C# and runs real-time with 30 ms per frame. The 10-fold classification is implemented in Matlab and takes only 33 ms in total. Both are tested on an Intel Core i7-3770K CPU 3.5 GHz with 8 GB RAM.

7.6 Conclusion

In this paper we introduced a new idea for sports type classification. Based on tracklets found by a Kalman filter we extract four simple, but robust, features. These are used for classification with a quadratic discriminant analysis. Using a total of 150 video sequences from five different sports types in a 10-fold cross validation we obtained a classification rate of 94.5 %. The result is better than what was previously obtained in [1], while this new approach is also more general; it doesn't depend on the position of the players or direction of play.

Due to privacy issues, we used thermal imaging only. However, the classification approach presented is applicable for other image modalities. Only the detection step should be substituted with a different method, which could be a HOG detector or another general person detector.

The proposed method is independent of the type of arena and it is expected that it could easily be extended to outdoor arenas as well. With the current set-up where the entire arena is monitored from a far-view, the level of details available for each person is limited. In a future perspective higher resolution imaging devices is expected to be available, enabling a more fine-grained analysis of individual people, such as pose and motion of each body part.

Acknowledgments

This project is funded by *Nordea-fonden* and *Lokale- og Anlægsfonden*, Denmark. We would also like to thank Aalborg Municipality for support and for providing access to their sports arenas.

References

[1] R. Gade and T. Moeslund, "Sports type classification using signature heatmaps," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2013.

[2] C. Krishna Mohan and B. Yegnanarayana, "Classification of sport videos using edge-based features and autoassociative neural network models," *Signal, Image and Video Processing*, vol. 4, pp. 61–73, 2010.

[3] Y. Yuan and C. Wan, "The application of edge feature in automatic sports genre classification," in *IEEE Conference on Cybernetics and Intelligent Systems*, 2004.

[4] P. Mutchima and P. Sanguansat, "TF-RNF: A novel term weighting scheme for sports video classification," in *IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC)*, 2012.

[5] D.-H. Wang, Q. Tian, S. Gao, and W.-K. Sung, "News sports video shot classification with sports play field and motion features," in *International Conference on Image Processing (ICIP)*, 2004.

[6] J. Wang, C. Xu, and E. Chng, "Automatic sports video genre classification using Pseudo-2D-HMM," in *18th International Conference on Pattern Recognition (ICPR)*, 2006.

[7] X. Gibert, H. Li, and D. Doermann, "Sports video classification using HMMS," in *International Conference on Multimedia and Expo (ICME)*, 2003.

[8] J. Y. Lee and W. Hoff, "Activity identification utilizing data mining techniques," in *IEEE Workshop on Motion and Video Computing (WMVC)*, Feb 2007.

[9] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[10] R. Gade, A. Jørgensen, and T. Moeslund, "Long-term occupancy analysis using graph-based optimisation in thermal imagery," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[11] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, 2001.

Part IV

Tracking sports players

Intuitively, when humans observe an object, they will visually follow and "track" the motion of the object. Thereby, information about the activities or behaviour of the object can automatically be inferred by the human brain. How to copy this ability by computers is being researched within a wide variety of fields.

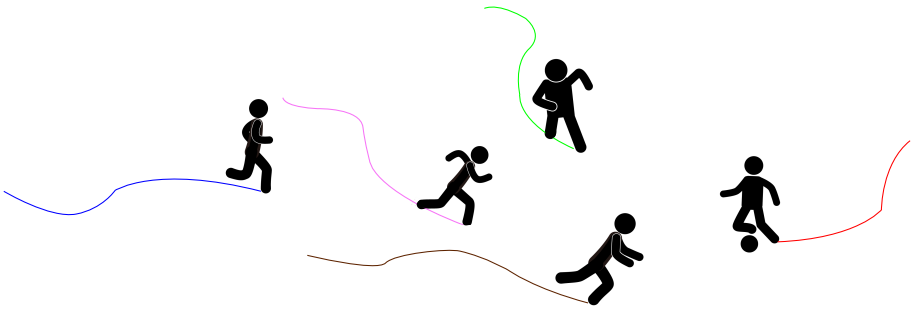
In sports analysis, tracking of players is the crucial first step in order to analyse both individual and team-based performance by extracting metrics, such as speed, distance, acceleration, direction, etc. Performance analysis are often seen presented in TV transmissions of sports games and used by coaches for improving the performance of athletes. However, most commercial tracking systems today are only semi-automatic and require skilled operators to guide the tracking of individual players.

The first chapter in this part deals with multi-target tracking of sports players applied to thermal video. In the second chapter we combine the work presented in chapter 5 on robust counting of people with a multi-target tracker on thermal video. The purpose is to improve the performance of a tracker by constraining the number of tracks produced. The last chapter in this part presents a method for improving multi-target tracking in RGB video, focusing on solving difficult situations of occlusions between people.

Chapter 9 consists of unpublished work in progress, while chapters 8 and 10 were originally published in a journal and in the proceedings of a workshop:

Rikke Gade and Thomas B. Moeslund, “Thermal Tracking of Sports Players,” *Sensors*, vol. 14, no. 8 pp. 13679–13691, July 2014.

Anton Milan, Rikke Gade, Anthony Dick, Thomas B. Moeslund and Ian Reid, “Improving Global Multi-target Tracking with Local Updates,” *ECCV workshop on Visual Surveillance and Re-Identification*, September 2014.



Chapter 8

Thermal Tracking of Sports Players

Rikke Gade and Thomas B. Moeslund

The paper has been published in
Sensors - special issue on Detection and Tracking of Targets in Forward-Looking InfraRed (FLIR) Imagery, Vol. 14(8), pp. 13679–13691, July 2014.

© 2014 MDPI
The layout has been revised.

Abstract

We present here a real-time tracking algorithm for thermal video from a sports game. Robust detection of people includes routines for handling occlusions and noise before tracking each detected person with a Kalman filter. This online tracking algorithm is compared with a state-of-the-art offline multi-target tracking algorithm. Experiments are performed on a manually annotated 2-minutes video sequence of a real soccer game. The Kalman filter shows a very promising result on this rather challenging sequence with a tracking accuracy above 70 % and is superior compared with the offline tracking approach. Furthermore, the combined detection and tracking algorithm runs in real time at 33 fps, even with large image sizes of 1920×480 *pixels.*

8.1 Introduction

Traditionally, visual cameras, capturing RGB or greyscale images, have been the obvious choice of sensor in surveillance applications. However, in dark environments, this sensor has serious limitations, if capturing anything at all. This is one of the reasons that other types of sensors are now taken into consideration. One of these sensors is the thermal camera, which has recently become available for commercial and academic purposes, although originally developed for military purposes [1]. The light-independent nature of this sensor makes it highly suitable for detection and tracking of people in challenging environments. Privacy has also become a big issue, as the number of surveillance cameras have increased rapidly. For video recording in sensitive locations, thermal imaging might be a good option to cover the identity of the people observed, in some applications it might even be the only legal video modality. However, like any other sensor type, the thermal sensor has both strengths and weaknesses, which are discussed in the survey on thermal cameras and applications [1]. One way of overcoming some of these limitations is to combine different sensors in a multi-modal system [2].

The visual and thermal sensors complement each other very well. Temperature and colour information are independent, and besides adding extra information on the scene each sensor might be able to detect targets in situations where the other sensor completely fails. However, registration and fusion of the two image modalities can be challenging, since there is not necessarily any relation between brightness level in the different spectra. Generally, three types of fusion algorithms exist; fusion on pixel level, feature level, or decision level. Several proposed fusion algorithms are summarised in the survey [1].

It is clear that multi-modal detection and tracking systems have several advantages for robust performance in changing environments, which is also shown in recent papers on tracking using thermal-visible sensors [3, 4]. The drawbacks of these fused systems primarily relates to the fusion part, which requires an additional fusion algorithm that might be expensive in time and

computations. Furthermore, when applying a visual sensor, the possibility of identification and recognition of people exists, causing privacy issues that must be considered for each application.

A direct comparison of tracking performance in multi-modal images versus purely thermal images in different environments would be interesting, but this is out of scope for this paper. Here we choose to take another step towards privacy-preserving systems and work with thermal data only. While tracking people in RGB and greyscale images has been and is still being extensively researched [5, 6], the research in tracking in thermal images is still rather limited. Therefore, in this paper we wish to explore the possibility of applying tracking algorithms in the thermal image modality.

8.1.1 Related Work

Two distinct types of thermal tracking of humans exist. One is tracking of human faces, which requires high spatial resolution and good quality images to detect and track facial features [7–9]. The other direction, which we will focus on, is tracking of whole-body individual people in surveillance-like settings. In this type of applications the spatial resolution is normally low and the appearance of people is very similar. We cannot rely on having enough unique features for distinguishing people from each other, and we must look for tracking methods using only anonymous position data.

For tracking in traditional RGB or greyscale video, the tracking-by-detection approach has recently become very popular [10–12]. The classifier is either based on a pre-trained model, e.g., a pedestrian model, or it can be a model-free tracker initialised by a single frame, learning the model online. The advantage of online learning is the ability to update the classifier, as the target may change appearance over time. In order to apply this approach for multi-target tracking, the targets should be distinguishable from each other. This is a general problem in thermal images, the appearance information is very sparse, as no colour, texture, *etc.*, are sensed by the camera.

Other approaches focus on constructing trajectories from "anonymous" position detections. Both online (recursive) and offline (batch optimisation) approaches has proven to be successful. Online approaches cover the popular Kalman filter [13] and particle filters [14, 15]. The methods are recursive, processing each frame as soon as it is obtained, and assigning the detection to a trajectory. Offline methods often focus on reconstructing the trajectories by optimising an objective function. Examples are presented in [16] by posing the problem as an integer linear program and solving it by LP-relaxation, or in [17] solving it with the k-shortest path algorithm.

Tracking in thermal video has often been applied in real-time applications for pedestrian tracking or people tracking for robot-based systems. Fast online approaches have therefore been preferred, such as the particle filter [18, 19] and the Kalman filter [20, 21].

While most works on tracking people in thermal images have focused on

pedestrians with low velocity and highly predictable motion, we apply tracking to real sports video, captured in a public sports arena. It is highly desired to track sports players in order to analyse the activities and performance of both teams and individuals, as well as provide statistics for both internal and commercial use. However, sports video is particularly challenging due to a high degree of physical interaction, as well as abrupt and erratic motion.

Figure 8.1 shows an example frame from the video used for testing. The video is captured with three cameras in order to cover the entire field of 20 m $\times$ 40 m. The images are rectified and stitched per frame to images of 1920×480 pixels.

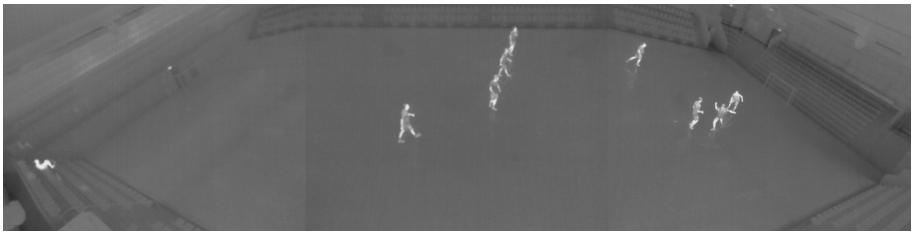


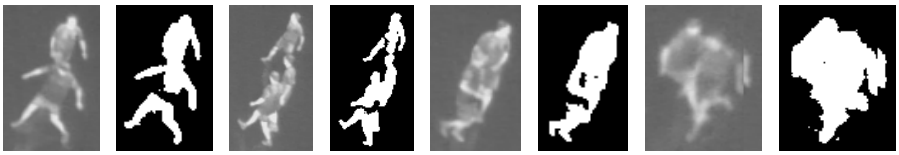
Fig. 8.1: Example of a frame from the thermal sports video.

This paper will investigate the applicability and performance of two different tracking approaches on thermal data. First, we design an algorithm based on the Kalman filter. Then, we test a publicly available state-of-the-art multi-target tracking algorithm [22]. The algorithms are evaluated on a 2 min manually annotated dataset from an indoor soccer game.

8.2 Detection

Detecting people in thermal images may seem simple, due to an often higher temperature of people compared with the surroundings. In this work we focus on indoor environments, more specifically a sports arena. This scene is quite simple in terms of a plain background with relatively stable temperature. Hence, people can often be segmented from the background by only thresholding the image. The challenges occur in the process of converting the binary foreground objects into individual people. In the ideal cases each blob is simply considered as one person. However, when people interact with each other, they overlap in the image and cause occlusions, resulting in blobs containing more than one person. The appearance of people in thermal images is most often as simple as grey blobs, making it impossible to robustly find the outline of individual people in overlaps. Figure 8.2 shows four examples of occlusions and the corresponding binarised images.

While full or severe occlusions (like Figure 8.2(e)) cannot be solved by detection on frame basis, we aim to solve situations where people are only partly occluded and can be split into single person. Likewise, we want to



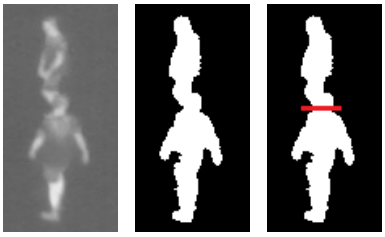
(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 8.2: Examples of occlusions between people. For each example the corresponding binarised image is shown, found by automatic thresholding.

detect only one person even when it has been split into several blobs during thresholding. We implement three rather simple but effective routines aiming at splitting or connecting the blobs into single person. These routines are described in the following sections.

8.2.1 Split Tall Blobs

People standing behind each other, seen from the camera, might be detected as one blob containing more than one person. In order to split these blobs into single detection we here adapt the method from [23]. First, it must be detected when the blob is too tall to contain only one person. If the blob has a pixel height that corresponds to more than a maximum height at the given position, found by an initialising calibration, the algorithm should try to split the blob horizontally. The point to split from is found by analysing the convex hull and finding the convexity defects of the blob. Of all the defect points, the point with the largest depth and a given maximum absolute gradient should be selected, meaning that only defects coming from the side will be considered, discarding, e.g., a point between the legs. Figure 8.3 shows an example of how a tall blob containing two people will be split.



(a) (b) (c)

Fig. 8.3: Example of how a tall blob containing two people will be split into two.

8.2.2 Split Wide Blobs

Groups of people standing next to each other might be found as one large blob. To identify which blobs contain more than one person, the height/width ratio and the perimeter are considered, as done in [23]. If the criteria are satisfied, the algorithm should try to split the blob. For this type of occlusion, it is often possible to see the head of each person, and split the blob based on the head positions. Since the head is narrower than the body, people can be separated by splitting vertically from the minimum points of the upper edge of a blob. These points can be found by analysing the convex hull and finding the convexity defects of the blob. Figure 8.4 shows an example of how a wide blob containing two people will be split.

(a) (b) (c)

Fig. 8.4: Example of how a wide blob containing two people will be split into two.

8.2.3 Connect Blobs

One person can often be split into several blobs during thresholding if some areas of the body appear colder, e.g., due to loose or several layers of clothing. In order to merge these parts into only one detected person, we consider each binary blob a candidate, and generate a rectangle of standard height at the given position (calculated during calibration) and the width being one third of the height. For each rectangle we evaluate the ratio of foreground (white) pixels. If the ratio of white pixels is below 15%, the blob is discarded, otherwise the candidate is added for further processing. The second step is to check if the candidate rectangles overlap significantly, hence probably belonging to the same person. If two rectangles overlap by more than 45%, only the candidate with highest ratio of white pixels is kept as a true detection. These threshold values are chosen experimentally by evaluating 340 positive samples and 250 negative samples. Figure 8.5 illustrates this situation, where one person has been split into three blobs.

The ultimate goal for the detection algorithm is to detect each person, and nothing else, in each frame. However, with a side-view camera angle and a number of people interacting, missing detections and noise must be considered when using the detections as input for the tracking algorithms described next.

(a) (b) (c)

Fig. 8.5: Example of a person that is split into three blobs by thresholding. Three overlapping candidates are evaluated (green, blue and yellow rectangles). Only the green candidate will be kept, because it has the highest ratio of white pixels.

8.3 Tracking

8.3.1 Kalman Filter

The Kalman filter, introduced in the early 1960s, is a now well-known algorithm used in a wide range of signal processing applications. The recursive algorithm filters noisy measurements by predicting the next step from previous state and use the new measurement as feedback for updating the estimate. The Kalman filter estimates the state x of a discrete-time controlled process controlled by the linear stochastic difference equation [24]:

$$x_k = Ax_{k-1} + Bu_{k-1} + w_{k-1} \tag{8.1}$$

with a measurement z:

$$z_k = Hx_k + v_k \tag{8.2}$$

where w_k and v_k are random variables representing the process and measurement noise, respectively. The matrix A is the transition matrix that relates the state x at the previous time step $k-1$ to the state at the current step k. The matrix B relates the control input u_{k-1} to the state x_k (the control input is optional, and this term is often discarded). The matrix H relates the state x_k to the measurement z_k.

Figure 8.6 illustrates the procedure of the Kalman filter, shifting between predicting the next step from the previous state and correcting the state using a new observed measurement.

Using the Kalman filter for tracking an object in 2D, the state x consists of four dynamic variables; x-position, y-position, x-velocity and y-velocity. The measurement z represents the observed x- and y-positions for each frame.

When implementing a Kalman filter, the measurement noise covariance R and the process noise covariance Q must be tuned. R represents the measurement noise variance, meaning that a high value will tell the system to rely less on the measurements and vice versa.

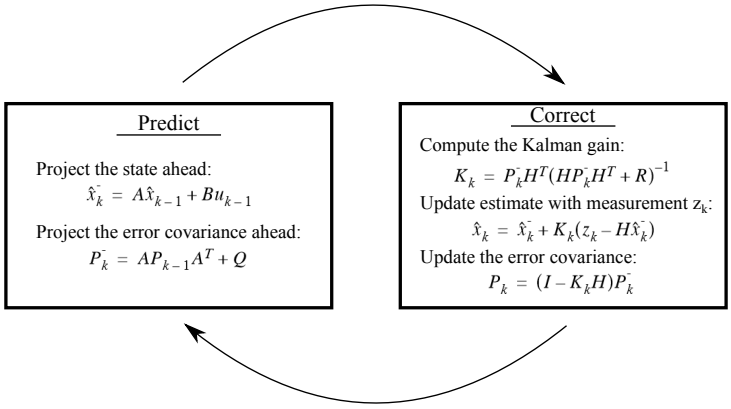


Fig. 8.6: Procedure of the Kalman filter.

For more details on the Kalman filter, we refer to the introduction in [24] or the original paper [13].

8.3.2 Multi-Target Data Association

Each Kalman filter maintains only the estimated state of one object. In order to keep track of several targets simultaneously, the association between detections and Kalman filters must be handled explicitly. For each frame, a list of detections are obtained as described in Section 8.2. Each existing Kalman filter is then assigned the nearest detection, within a given distance threshold th. For each detection that is not assigned to a Kalman filter, a new track is started, by creating a new Kalman filter. Kalman filters that have no assigned detections will be continued based on the predicted new positions. After a given time period without detections, experimentally set to 10 frames, the track will be terminated.

8.3.3 Tracking by Continuous Energy Minimization (CEM)

The choice of tracking based on the Kalman filter leaves no possibility for connecting broken tracks, as it is a purely recursive approach. This possibility of optimising both forward and backward in time is instead exploited in offline algorithms based on batch optimisation. We will here test one of these algorithms, using code available online. This algorithm minimises an energy function of five terms [22]:

$$E(X) = E_{obs} + \alpha E_{dyn} + \beta E_{exc} + \gamma E_{per} + \delta E_{reg} \tag{8.3}$$

E_{obs} represents the likelihood of object presence, determined by the object detector. E_{dyn} is the dynamic model, using a constant velocity model. E_{exc} is a mutual exclusion term, introducing the physical constraint that two objects

cannot be present at the same space simultaneously. The target persistence term E_{per} penalises trajectories with start or end points far from the image border. The last term, E_{reg}, is a regularisation term that favours fewer targets and longer trajectories.

Given the set of detections for all frames, this algorithm will try to minimise the energy function (8.3) by growing, shrinking, splitting, merging, adding or removing until either convergence or reaching the maximum number of iterations. For further details, see [22].

The set of detections are found as described in Section 8.2. Being the same detection algorithm used for both Kalman filter and CEM tracker, the tracking algorithms can be compared directly.

8.4 Experiments

In this section we test the tracking algorithm on a 2-minutes (3019 frames) video from an indoor soccer game. The frames are manually annotated using bbLabeler from Piotr's Image & Video MATLAB Toolbox [25]. All frames are annotated by the same person in order to ensure consistency.

8.4.1 Kalman Tracker

The Kalman filter tracker is implemented in C# using EMGU CV wrapper for the OpenCV library [26]. Through experiments the measurement noise covariance R has been tuned to 0.1 and the process noise covariance Q is tuned to 0.002 for position and 0.003 for velocity.

8.4.2 CEM Tracker

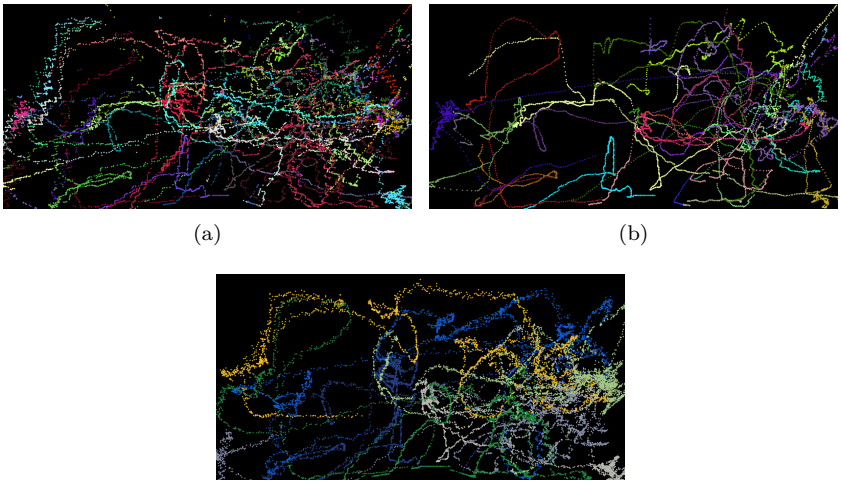
The CEM tracker is downloaded from the author's website[1]. We use the 2D tracking option, tracking in image coordinates. Default parameter values are used, except for three parameters: Target size is reduced to ImageWidth/200 (approx. 10 pixels) due to the relatively small object size in the test video. The maximum number of global iterations is varied between 15, 30 and 60 iterations, along with the maximum number of iterations for each gradient descent, which is varied between 30, 60 and 120 iterations.

8.4.3 Results

The trajectories found by the Kalman tracker, CEM tracker and manually annotated trajectories, respectively, are plotted in Figure 8.7. The trajectories are plotted in world coordinates, thus each image represents the sports field seen from above. Each new identity found by the tracker is plotted in a new

[1] http://www.milanton.de/contracking/index.html

colour assigned randomly. The figure shows that while the trajectories found by the CEM tracker is longer and smoother, the Kalman tracker produces more tracks, which are also very close to the ground truth. We evaluate the track-



(a) (b)

(c)

Fig. 8.7: Trajectories plot in world coordinates with each identity assigned a random colour. (**a**) Trajectories found by Kalman tracker, (**b**) trajectories found by CEM tracker (60 epochs) and (**c**) manually annotated trajectories.

ing results using CLEAR MOT metrics [27], calculated by publicly available MATLAB code [28]. The results are measured by true positives (TP), false positives (FP), false negatives (FN), ID switches and the two combined quality measures: multiple object tracking precision (MOTP) and multiple object tracking accuracy (MOTA):

$$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_t c_t} \tag{8.4}$$

where d_t^i is the distance between the object o_i and its corresponding hypothesis. c_t is the number of matches found for time t. Hence, MOTP is the total error in estimated position for matched object–hypothesis pairs over all frames, averaged by the total number of matches made.

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDS_t)}{\sum_t g_t} \tag{8.5}$$

where FN_t, FP_t and IDS_t are the number of false negatives, false positives and ID switches, respectively, for time t, while g_t is the true number of objects at time t.

The results of the Kalman tracker and the CEM tracker with three different numbers of maximum iterations are presented in Table 8.1. The result for

	TP	FP	FN	ID sw.	MOTP	MOTA	Tot. tr. length	#ID's
KF	80.22%	9.86%	18.86%	219	0.75	70.36%	2506.78m	218
CEM-15	11.61%	27.38%	88.14%	60	0.58	-15.77%	933.66m	37
CEM-30	17.07%	33.19%	82.72%	51	0.59	-16.11%	1100.69m	31
CEM-60	18.14%	38.06%	81.60%	60	0.60	-19.91%	1228.14m	33

Table 8.1: Tracking results for Kalman filter and continuous energy minimization (CEM) algorithms.

Kalman filtering is very good, considering the complexity of the data. The true positive rate exceeds 80% and the accuracy (MOTA) is 70.36%. For the CEM tracker these numbers are significantly lower, the best true positive rate is 18.14% obtained after 60 epochs. This implies a high false negative rate of 81.6%, but also a high false positive rate of 38.06%. The resulting MOTA ends up being negative. The results are clearly related to the total track length; the Kalman filter constructs more than twice the total length of tracks, which is closer to the total length of ground truth tracks of 3241.52 m.

8.4.4 Processing Time

The processing time, calculated for 3019 frames of video containing 8 people, is for the MATLAB implementation of the CEM tracker (excluding detection) with 15 epochs: 6.03 min (0.12 s per frame), 30 epochs: 8.75 min (0.17 s per frame), 60 epochs: 16.34 min (0.32 s per frame). For the C# implementation of Kalman filter tracking (with integrated detection) the processing time is only 1.55 min (0.03 s per frame).

Both methods are tested on an Intel Core i7-3770K CPU 3.5 GHz with 8 GB RAM.

8.5 Discussion

We have tested the CEM tracker with three different numbers of maximum iterations, in order to investigate whether more iterations would allow the algorithm to reach a better estimate. From 15 to 30 epochs we observe clear improvements, from a true positive rate of 11.61% to 17.07% and the false negative rate decreasing accordingly. The false positive rate increases from 27.38% to 33.19%, though. Increasing the maximum number of iterations from 30 to 60 gives only a small improvement in true positive rate from 17.07% to 18.14%, while the false positive rate increases from 33.19% to 38.06%. This indicates that further iterations will not improve the accuracy.

Given that the CEM tracker is an offline algorithm, processing a batch of frames, it is able to run the optimisation both forward and backward in time. That makes it more likely to connect broken trajectories compared with the Kalman tracker, which is recursive and needs to start a new trajectory if it loses one. As expected, this is observed as more identity switches by the Kalman tracker (219 switches) compared with the CEM tracker (51–60 switches). It is

also reflected in the mean length of each trajectory; for the Kalman tracker the mean length is 11.5 m, compared with 25.2–37.2 m for the CEM tracker.

The processing time of the two algorithms indicates another big difference between online and offline approaches. The Kalman filter is well-suited for real-time applications with a processing time of only 0.03 s per frame including both detection and tracking. For the CEM tracker the detections must be saved for the full batch of frames before starting to construct trajectories. The processing time is then 0.12–0.32 s per frame, depending on the number of iterations. Furthermore, the processing time might increase significantly with the number of targets.

Both tracking algorithms are independent of the type of detection algorithm, making it possible to apply tracking to a wide range of applications. In this work we demonstrated the approach on a video from an indoor sports arena, but it could be applied directly in any scene where the human temperature is different from the background, including outdoor scenes. The performance depends on the quality of detections. In order to significantly reduce the occlusions between people, the camera could be mounted above the scene, capturing a top-view instead of the side-view shown in Figure 8.1.

8.6 Conclusion

We have presented an online multi-target tracking algorithm based on the Kalman filter and compared with a state-of-the-art offline multi-target tracking algorithm. In terms of accuracy the Kalman tracker is far superior in this application and constructs more than twice the total length of tracks. The drawback of this online approach is the number of split tracks and identity switches. Depending on the application and importance of identity, a post-processing method could be applied in order to optimise and connect trajectories.

Acknowledgements

This project is funded by *Nordea-fonden* and *Lokale-og Anlægsfonden*, Denmark. We would also like to thank Aalborg Municipality for support and for providing access to their sports arenas.

References

[1] R. Gade and T. Moeslund, "Thermal cameras and applications: a survey," *Machine Vision and Applications*, vol. 25, no. 1, pp. 245–262, 2014.

[2] Z. Zhu and T. S. Huang, Eds., *Multimodal Surveillance: Sensors, Algorithms and Systems.* Artech House Publisher, 2007.

[3] M. Airouche, L. Bentabet, M. Zelmat, and G. Gao, "Pedestrian tracking using color, thermal and location cue measurements: a DSmT-based framework," *Machine Vision and Applications*, vol. 23, pp. 999–1010, 2012.

[4] A. Torabi, G. Massé, and G.-A. Bilodeau, "An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications," *Computer Vision and Image Understanding*, vol. 116, no. 2, pp. 210–221, 2012.

[5] J. Watada, Z. Musa, L. Jain, and J. Fulcher, "Human tracking: A state-of-art survey," in *Knowledge-Based and Intelligent Information and Engineering Systems*, ser. Lecture Notes in Computer Science, R. Setchi, I. Jordanov, R. Howlett, and L. Jain, Eds. Springer Berlin Heidelberg, 2010, vol. 6277, pp. 454–463.

[6] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013, pp. 2411–2418.

[7] A. Alkali, R. Saatchi, H. Elphick, and D. Burke, "Facial tracking in thermal images for real-time noncontact respiration rate monitoring," in *European Modelling Symposium (EMS)*, Nov 2013, pp. 265–270.

[8] F. AL-Khalidi, R. Saatchi, D. Burke, and H. Elphick, "Tracking human face features in thermal images for respiration monitoring," in *IEEE/ACS International Conference on Computer Systems and Applications (AICCSA)*, May 2010, pp. 1–6.

[9] W. Lee, K. Jung, Y. Kim, G. Lee, and C. Park, "Implementation of face tracking system for non-contact respiration monitoring," in *Lecture Notes in Information Technology vol. 14*, 2012, pp. 160–163.

[10] Z. Kalal, J. Matas, and K. Mikolajczyk, "P-n learning: Bootstrapping binary classifiers by structural constraints," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2010, pp. 49–56.

[11] S. Hare, A. Saffari, and P. Torr, "Struck: Structured output tracking with kernels," in *IEEE International Conference on Computer Vision (ICCV)*, Nov 2011, pp. 263–270.

[12] J. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Computer Vision – ECCV 2012*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, vol. 7575, pp. 702–715.

[13] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[14] J. Vermaak, A. Doucet, and P. Perez, "Maintaining multimodality through mixture tracking," in *Ninth IEEE International Conference on Computer Vision, Proceedings*, Oct 2003, pp. 1110–1116 vol.2.

[15] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Robust tracking-by-detection using a detector confidence particle filter," in *IEEE 12th International Conference on Computer Vision*, Sept 2009, pp. 1515–1522.

[16] J. Berclaz, F. Fleuret, and P. Fua, "Multiple object tracking using flow linear programming," in *Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter)*, Dec 2009, pp. 1–8.

[17] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1806–1819, Sept 2011.

[18] P. Skoglar, U. Orguner, D. Törnqvist, and F. Gustafsson, "Pedestrian tracking with an infrared sensor using road network information," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, pp. 1–18, 2012.

[19] A. Treptow, G. Cielniak, and T. Duckett, "Real-time people tracking for mobile robots using thermal vision," *Robotics and Autonomous Systems*, vol. 54, no. 9, pp. 729–739, 2006.

[20] K. Jüngling and M. Arens, "Local feature based person detection and tracking beyond the visible spectrum," in *Machine Vision Beyond Visible Spectrum*, ser. Augmented Vision and Reality. Berlin: Springer-Verlag, 2011, vol. 1, pp. 3–32.

[21] S. Lee, G. Shah, A. Bhattacharya, and Y. Motai, "Human tracking with an infrared camera using a curve matching framework," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, pp. 1–15, 2012.

[22] A. Andriyenko and K. Schindler, "Multi-target tracking by continuous energy minimization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011, pp. 1265–1272.

[23] R. Gade, A. Jørgensen, and T. B. Moeslund, "Occupancy analysis of sports arenas using thermal imaging," in *Proceedings of the International Conference on Computer Vision and Applications*, 2012.

[24] G. Welch and G. Bishop, "An introduction to the kalman filter," Chapel Hill, NC, USA, Tech. Rep., 1995.

[25] P. Dollár, "Piotr's Image and Video Matlab Toolbox (PMT)," http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html.

[26] EMGU CV, "Documentation," http://www.emgu.com/wiki/index.php/Documentation.

[27] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, p. 246309, 2008.

[28] A. D. Bagdanov, A. D. Bimbo, F. Dini, G. Lisanti, and I. Masi, "Posterity logging of face imagery for video surveillance," *IEEE Multimedia*, vol. 19, no. 4, pp. 48–59, 2012.

Chapter 9

Constrained Multi-target Tracking for Thermal Imaging

Rikke Gade

The paper is unpublished work in progress.

The layout has been revised.

Abstract

The ability to track multiple people simultaneously is useful in many applications, e.g. sports analysis and analysis of human behaviour in urban spaces. However, it is still a challenging task with no general solution. Especially for thermal data, robust tracking solutions is still lacking. In this work we aim to guide a global tracker with estimated knowledge on periods with a stable number of people in the scene. Using the counting algorithm presented in chapter 5 and combining it with a global offline tracker, we show clear improvement on a test set of four thermal video sequences of 30 seconds each. Compared to the original offline tracking algorithm, we obtain improvements from 1-16% in the accuracy.

9.1 Introduction

After several years of research in tracking, consistent tracking of multiple people is still very challenging. Human motion can be erratic, and interactions between people complicates the task a lot. In most cases video is captured from either eye-level or high-angle, but rarely from the optimal perpendicular top-view. Therefore, occlusions must be considered, and solving these ambiguous situations require some type of data association.

In this work we apply tracking to thermal video from both indoor and outdoor scenes. Thermal video has a great potential in surveillance and any application where privacy is an issue, such as analysis of behaviour in public spaces and public buildings. The number of computer vision methods tested on thermal video is still relatively low, leaving the applicability and performance on this type of data unknown.

Fig. 9.1: Example of an outdoor courtyard captured by a thermal camera.

The properties of thermal video in far-view, surveillance-like settings, will

result in mostly identical appearance of people. The lack of distinct appearance information makes re-identification after full occlusions impossible. We must rely on motion, even though some activities, such as sport, may include erratic motion.

In this work we introduce a method to construct more reliable tracks. We include an estimation of the occupancy of the observed scene and the duration of stable periods without people entering or leaving the scene. We show that including this constraint on stable periods improves the overall tracking performance. For multi-target tracking offline approaches have become increasingly popular, due to their superior accuracy. Compared to online (recursive) approaches, offline methods have great advances in that they optimise trajectories over batches of frames. However, big challenges still remain in many applications, due to noise and ambiguities. From a (probably noisy) set of detections, the algorithm must construct an unknown number of trajectories. This task causes ambiguities and thereby errors or inaccuracies. In this work we aim to take advantage of automatic preprocessing, which can help limiting the problem by estimating how many people are present in the scene. The next section will provide an overview of the proposed method.

9.2 Overview

We propose a tracking algorithm which runs in two main iterations. In the first iteration we estimate time periods which can be characterised as stable periods, without people leaving or entering the scene, as well as the probability of a given number of people present during that period. In the second iteration the result is fed to a tracking algorithm in order to constrain the number of trajectories produced during each of the stable periods. The algorithms is illustrated in figure 9.2. The hatched areas in the top row represents stable periods, from which the estimated number is fed to the tracking algorithm in the second iteration. During non-stable periods no constrain is added, leaving the tracker to try to connect the paths using the original algorithm.

9.3 Counting People

In most applications the recorded scene consists of an area where people move around freely, and some possible entrance/exit areas. These entrance/exit areas might only be the edge of the image, or there might be doors in the scene. Assuming that people are not constantly moving in and out of the scene, the number of people observed in the scene will stay constant during a time period.

An estimation of this number can be calculated using the approach presented in [1], which will briefly be described here.

Firstly, we must try to detect all people in each frame. For segmentation purposes background subtraction is applied, followed by automatic threshold-

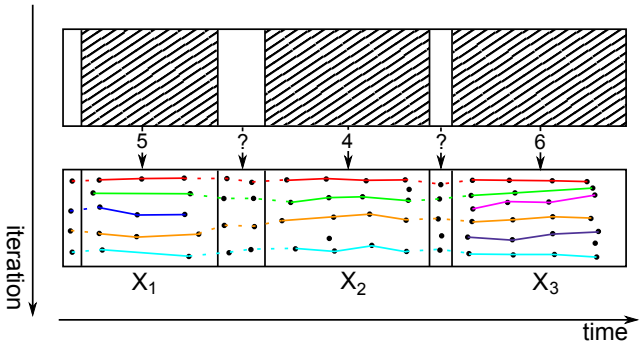


Fig. 9.2: Illustration of the proposed method. During first iteration, stable periods of the video sequence are identified and a number of people present are estimated. This is used as input for the second iteration, in which trajectories are constructed and optimised.

ing. The resulting binary objects are then examined, and optionally split vertically or horizontal if they are probable to contain more than one person. This procedure is described in detail in [2].

An uncertainty is still related to each binary object, for it being a true person detected or not. The probability of being a true detection is related to the ratio of white pixels within the bounding box and the ratio of white pixels observed on the edge of the bounding box. Experimentally it is found that the highest probability is at a ratio of white pixels between 30 and 60%. Furthermore, less than 50% of the edge is allowed to be white. The weighting related to the ratio of white pixels in the rectangle (r_r) and the ratio of white pixels on the perimeter (r_p) are described in eq. 9.1:

$$w_p(i) = \begin{cases} 0, & \text{if } r_p > 50\% \parallel r_r < 20\% \\ 0.8, & \text{if } r_r > 70\% \\ 0.9, & \text{if } r_r < 30\% \parallel r_r > 60\% \\ 1, & \text{otherwise} \end{cases} \tag{9.1}$$

The weighting of each detection is combined with a weight describing the uncertainty for each frame, caused by occlusions and clutter. Each frame counting is weighted like this:

$$w_f = \alpha \cdot \prod_{i=1}^{n} w_p(i) + \beta \cdot w_s \tag{9.2}$$

where n is the number of people, $w_p(i)$ is the probability of person i being a true detection (see equation 9.1), and w_s is a weight that decreases with the number of splits performed, indicating how cluttered the scene is. α and β are the weighting of each part and should sum to one. The observed number in a frame will be added to a histogram with the weight w_f, and after a stable period has ended, the histogram will be scaled to an accumulated sum of 1. The circles in figure 9.3 illustrate the weighted histogram for each period.

In order to detect stable periods of a video sequence, we must detect when people are close to the border of the scene (and if applicable any other entrance/exit in the scene). These periods should be flagged and observed for people leaving or entering the scene. This is done by applying local tracking on people close to the border. All other periods are marked as stable periods, and should contain the same number of people until next period of border activity. Estimating the number of people are done by frame-based detection succeeded by an graph optimisation algorithm, based on Dijkstra's algorithm [3]. The graph optimisation interprets the stable periods as nodes and transitions (people leaving or entering the scene) as edges. All nodes and edges have a weight factor based on the detection and tracking results.

Figure 9.3 illustrates the graph approach. For each stable period the number of people are represented by circles where darker colour is higher weight. The lines between two stable periods represent the transitions, also coloured darker for higher weight. The path through this graph is optimised to the highest total weight.

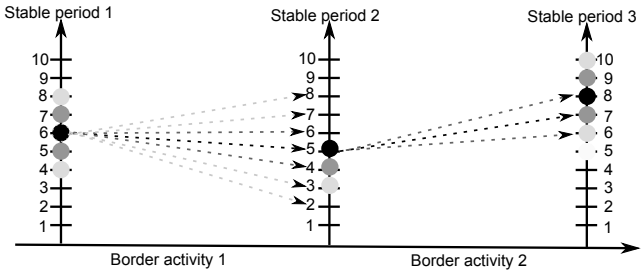


Fig. 9.3: Example of a simple graph. Dark nodes and edges have the highest weight. Edges exist between all nodes in two consecutive periods, but to simplify the illustration the edges with lowest weight are not drawn.

9.4 Tracking by Energy Minimization

As a starting point for the offline tracking algorithm we use the algorithm proposed by Andriyenko and Schindler [4], which has shown very good results on real-world datasets. It has publicly available source code[1], which we can use for further testing. The aim of this method is to find the optimal solution for multi-target tracking over an entire video sequence, given a set of coordinates of all targets at all times. The core part of this algorithm is to minimise the following global energy function:

$$E(x) = E_{obs} + \alpha E_{dyn} + \beta E_{exc} + \gamma E_{per} + \delta E_{reg} \tag{9.3}$$

[1] http://www.milanton.de/contracking/

E_{obs} represents the likelihood of object presence, determined by the object detector. E_{dyn} is the dynamic model, using a constant velocity model. E_{exc} is a mutual exclusion term, introducing the physical constraint that two objects cannot be present at the same space simultaneously. The target persistence term E_{per} penalises trajectories with start or end points far from the image border. The last term E_{reg} is a regularisation term that favours fewer targets and longer trajectories. As E_{reg} is exactly a term considering the number of target, we will investigate if the constraint could be integrated in this term. The original E_{reg} term proposed in [4] is defined as follows:

$$E_{reg}(X) = N + \sum_{i=1}^{N} \frac{1}{F(i)} \tag{9.4}$$

where F is the temporal length of trajectory i in frames and N is the total number of trajectories. The first part of the equation is the total number of trajectories, inferring that the energy directly increases with this number. The second part is the sum of the inverse length of all trajectories, hence, in the minimisation process it will favour long trajectories.

9.5 Constraining the Tracking Algorithm

We aim to constrain the tracking algorithm, to construct n trajectories, where n is the number with highest probability, estimated by the counting algorithm described in section 9.3. Two relevant parameters can intuitively be formulated: the number of targets tracked per frame, and the total number of trajectories in each stable period. Ideally, since we are only concerned about stable periods, the total number of trajectories should correspond to the number of targets tracked in each frame. However, if the trajectory of one person is fragmented into shorter tracks, the total number of trajectories will increase while the correct number of targets can still be tracked in every frame. Likewise, if the target is lost during the sequence, the total number of trajectories might be correct, while some frames have less targets. Therefore, both measures might be valid parameters to include in the optimisation.

A and B represents how close the number of targets is to the estimated number, per frame and per stable period, respectively:

$$A = \frac{1}{F} \sum_{i=1}^{F} P(s(i), n(i)) \tag{9.5}$$

$$B = \frac{1}{S} \sum_{x=1}^{S} P(S(x), N(x)) \tag{9.6}$$

where $P(x)$ is the probability function for x number of targets. $n(i)$ is the number of targets existing in frame i, and $N(x)$ is the total number of trajectories

constructed per stable period x. F is the total number of frames.

Including the original two terms we now have four possible terms with following purposes:

1. Minimize number of targets
2. Maximize length of tracks
3. Constrain number of targets per frame
4. Constrain number of tracks per stable period

Since we now know the estimated number of people during each period, the original term 1 which is minimizing the number of targets is conflicting the purposes of terms 3 and 4, which add more specific constraints on the number. As a result, we discard term 1 and propose a new E_{reg} term including eq. 9.5 and 9.6:

$$E_{reg}(X) = \sum_{i=1}^{N} \frac{1}{F(i)} + w_1 \frac{1}{F} \sum_{i=1}^{F} P(s(i), n(i)) + w_2 \frac{1}{S} \sum_{x=1}^{S} P(S(x), N(x)) \tag{9.7}$$

A weight is added to each term, adjusting the influence from each term. These weights will be fitted during an optimisation process, described in section 9.6.

9.6 Evaluation

9.6.1 Datasets

To prove the robustness of our proposed method, we test on two different thermal datasets from one indoor and one outdoor environment. The main dataset we use for both test and training is captured in an indoor sports arena. In order to cover the entire field of 20×40 meters, three images are captured simultaneously and are stitched horizontal to a total image size of 1920×480 pixels. The dataset is captured during a soccer game with six to eight players at the court in all frames. Two minutes of video are captured and manually annotated for tracking. It is separated into four sequences of 30 seconds each in order to have a temporal window manageable for a global tracking algorithm. One sequence is used for training and the remaining three are used for testing. Figure 9.4 shows a frame from this soccer dataset.

The second dataset used is 30 seconds of video captured in an outdoor courtyard environment. A total of 14 different people walk through the courtyard, some of them in groups, causing difficult occlusions. The resolution of the images are 640×480 pixels. Figure 9.1 shows a frame from this dataset.

Fig. 9.4: A frame from the soccer dataset.

9.6.2 Weight parameters

The parameters of the original energy function, eq. 9.3, are kept to the default values for 2D tracking, given in the publicly available implementation: $\alpha = 1, \beta = 1, \gamma = 0.5, \delta = 1$.

The weight parameters w_1 and w_2 introduced in section 9.4 are here fitted experimentally in order to adjust the influence of each term. We use the 30 seconds training sequence, described in section 9.6.1. All combinations of the following parameter values are tested for w_1 and w_2: {0, 0.1, 1, 5, 10, 15, 20}.

The results seem to be most sensitive to w_2, where the accuracy is highest at $w_2 = 10$, while the accuracy varies less than 1 % with w_1. We fix $w_1 = 0.1$ which has a slightly higher accuracy.

9.6.3 Results

We compare the results of our method to the original formulation of the tracking algorithm presented in [4].

For evaluating the performance we use the multiple object tracking precision (MOTP) and multiple object tracking accuracy (MOTA) defined in the CLEAR MOT metrics [5]:

$$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_t c_t} \tag{9.8}$$

where d_t^i is the distance between the object o_i and its corresponding hypothesis. $c_t$is the number of matches found for time t. Hence, MOTP is the total error in estimated position for matched object-hypothesis pairs over all frames, averaged by the total number of matches made.

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDS_t)}{\sum_t g_t} \tag{9.9}$$

where FN_t, FP_t and IDS_t are the number of false negatives, false positives and ID switches, respectively, for time t, while g_t is the true number of objects at time t.

The results are presented in tables 9.1-9.4

	TP	FP	FN	ID switch	MOTP	MOTA
Original	74.05 %	11.8 %	25.53 %	25	19.99 cm	62.25 %
Ours	79.53 %	15.93 %	19.72 %	45	19.89 cm	**63.6** %

Table 9.1: Results - soccer sequence 1.

	TP	FP	FN	ID switch	MOTP	MOTA
Original	76.78 %	6.32 %	22.83 %	23	15.48 cm	70.46 %
Ours	84.32 %	8.52 %	15.04 %	38	15.47 cm	**75.80** %

Table 9.2: Results - soccer sequence 2.

It is clear for all sequences, that the number of true positives increases significantly, and the number of false negatives decreases. However, we see a small increase in false positives as well as the number of ID switches. The final results clearly show improvements on all sequences, with 1-16% increase in MOTA. The precision (MOTP) is unchanged for all sequences.

9.7 Conclusion

In this work we have shown how to combine a counting algorithm with an offline tracking algorithm, in order to constrain the number of tracks and improve the reliability. The preliminary results show good improvement on four sequences from both indoor and outdoor scenarios. As this is still work in progress, we will continue with more evaluations of the methods and comparisons to other work.

	TP	FP	FN	ID switch	MOTP	MOTA
Original	74.67 %	6.99 %	24.91 %	25	19.70 cm	67.69 %
Ours	80.49 %	8.19 %	18.92 %	35	19.79 cm	**72.31** %

Table 9.3: Results - soccer sequence 3.

	TP	FP	FN	ID switch	MOTP	MOTA
Original	48.58 %	12.42 %	51.24 %	10	19.94 cm	36.15 %
Ours	69.80 %	16.85 %	29.66 %	29	18.77 cm	**52.95** %

Table 9.4: Results - Courtyard sequence.

References

[1] R. Gade, A. Jørgensen, and T. Moeslund, "Long-term occupancy analysis using graph-based optimisation in thermal imagery," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[2] R. Gade, A. Jørgensen, and T. B. Moeslund, "Occupancy analysis of sports arenas using thermal imaging," in *Proceedings of the International Conference on Computer Vision and Applications*, feb. 2012.

[3] E. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[4] A. Andriyenko and K. Schindler, "Multi-target tracking by continuous energy minimization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011, pp. 1265–1272.

[5] K. Bernardin and R. Stiefelhagen, "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, p. 246309, 2008.

Chapter 10

Improving Global Multi-target Tracking with Local Updates

Anton Milan, Rikke Gade, Anthony Dick, Thomas B. Moeslund and Ian Reid

The paper has been published in
Proceedings of the European Conference on Computer Vision (ECCV) Workshops, September 2014.

© 2014 Springer
The layout has been revised.

Abstract

We propose a scheme to explicitly detect and resolve ambiguous situations in multiple target tracking. During periods of uncertainty, our method applies multiple local single target trackers to hypothesise short term tracks. These tracks are combined with the tracks obtained by a global multi-target tracker, if they result in a reduction in the global cost function. Since tracking failures typically arise when targets become occluded, we propose a local data association scheme to maintain the target identities in these situations. We demonstrate a reduction of up to 50% *in the global cost function, which in turn leads to superior performance on several challenging benchmark sequences. Additionally, we show tracking results in sports videos where poor video quality and frequent and severe occlusions between multiple players pose difficulties for state-of-the-art trackers.*

10.1 Introduction

Tracking multiple objects in a dynamic environment is crucial for visual scene understanding. Some of the most relevant applications for this task include driver assistance, visual surveillance, and sports analysis. The problem itself consists of localising each target in every single time instance *as well as* correctly maintaining each target's identity over time. This latter task is often referred to as *data association* and can be solved by existing methods as long as all targets remain sufficiently far apart from one another. However, challenges arise when several targets come close together causing intersecting or intertwined trajectories. In such situations, recovering each individual's identity has a combinatorial complexity in the number of tracks and measurements, and thus quickly becomes infeasible. In addition, the task is complicated further by noisy sensor data with imprecise localisation, false alarms, and missing measurements.

Most current approaches to multi-target tracking are based on *tracking by detection* [1–6]. Here, tracks are formed by linking detections obtained independently in each frame in a preprocessing step. This helps to avoid tracker drift, but usually depends on a pre-defined target model which is trained offline. When tracking by detection, more accurate results have been obtained by so-called *global* methods that consider a batch of several frames (or even an entire video sequence) jointly as opposed to determining the state based only on previous observations [7, 8]. The rationale here is that potential ambiguities may be resolved more easily once more evidence is acquired. However one must accept a delay in the output as a tradeoff for better accuracy.

Although tracking by detection approaches achieve state-of-the-art results, they struggle in those situations where the detector provides little to no evidence for the presence of a target. Detector failures may arise for numerous reasons, such as low image contrast, partial or complete occlusions, or abrupt

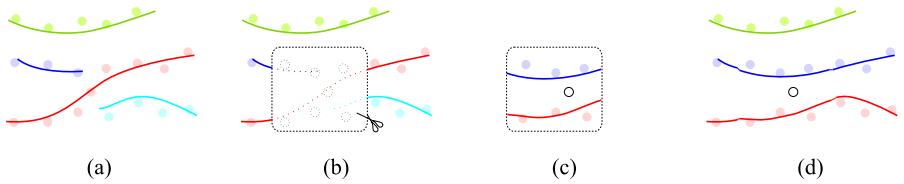


Fig. 10.1: Overview of our optimisation algorithm. Given a possibly erroneous solution *(a)*, we locate each error *(b)* and perform a local optimisation within its neighbourhood *(c)*. The newly obtained solution is inserted back into the original one if and only if it increases the overall likelihood considering all remaining frames and targets *(d)*.

and significant change in appearance due to lighting, posture, or object size. Even though short detection dropouts in certain, unambiguous areas can usually be bridged robustly by global optimisation techniques, correctly resolving data association remains challenging in cases where several targets merge on the image plane obstructing each other's line of sight. Long term occlusions or ambiguities are even more challenging, as the number of feasible association combinations increases with the time interval considered.

We propose to exploit the power of *model-free visual trackers* to 'untangle' tracks in such challenging situations (see Fig. 10.1 for an illustration). Model-free trackers do not rely on pre-existing detections, instead building an online model of target appearance based purely on an instance of the target appearing in a single frame. The performance of visual object trackers has increased dramatically in recent years [9] making them robust to appearance change and partial occlusion, which is a desirable property for solving the problem at hand. Moreover, we propose a strategy to integrate model-free visual object tracking into a multi-target tracking setting. Although visual trackers have, in one way or another, been previously used in combination with multiple target tracking [10–12], we present a rather different strategy to couple the two approaches.

In particular, our main contributions are as follows:

- We propose a scheme to explicitly detect challenging situations in multi-target tracking and address these in a way that builds on recent progress in both single and multi-target tracking.
- We apply model-free visual trackers to several targets simultaneously in order to resolve difficult situations locally.
- We integrate visual trackers into a multi-target tracking framework to find improved optima of the objective function by making local changes.
- We demonstrate the validity of our approach on particularly challenging sports videos.

We argue that our approach is able to drive the optimisation much quicker towards improved local minima leading to a substantial increase in performance

both visually and quantitatively. Our experiments show superior performance on several challenging benchmark sequences.

10.2 Related Work

The popularity of multi-target tracking in computer vision has increased dramatically in the recent past leading to a large amount of related literature. In this section we will concentrate on the most important work related mainly to offline multi-target tracking approaches. Despite their limitation of a delayed output, offline approaches to multi-target tracking have become increasingly popular due to their superior accuracy. The main difference to online (or recursive) approaches, such as Kalman filters [13, 14] or particle filters [7, 8] is that instead of processing each frame as soon as it is obtained, the optimisation of an objective function is performed on a batch of consecutive frames simultaneously. These methods are usually more robust at dealing with false positives or occlusions.

Offline multi-target tracking.

The main difference between approaches lies in the exact formulation of the objective function and its optimisation strategy. Jiang et al. [15] solve an integer linear program using LP-relaxation to obtain a (nearly) optimal solution. However, the number of targets in the scene needs to be fixed a-priori. Zhang et al. [1] reformulate the task as a network flow problem, which can be solved in polynomial time using min-cost flow algorithms. Occlusions are handled by inserting target hypotheses in a greedy fashion. Their approach served as a starting point for a similar strategy [16], which followed a greedy optimisation scheme and was thus much more efficient. Another globally optimal approach, which explicitly models merged measurements is presented in [17]. Individual tracks are however resolved using a simple shortest paths strategy, which may result in intersecting paths. More recently, Liu et al. [5] use a network-flow approach to recover long-term trajectories of sports players using context-aware motion models, while Butt and Collins [18] integrate high-order dynamic terms. A coupling of object detection and tracking has been proposed in [19, 20] with a quadratic and linear objective, respectively. Further formulations to solve for data association include graph-based approaches, such maximum weight independent set [21] set-cover [22] and generalised minimum clique graphs [4].

A slightly different way to solve the task is to concentrate on reconstructing trajectories rather than on data association and only implicitly handle the latter. A regularly discretised space allows one to pose the problem as an integer linear program, which is solved to global optimality by LP-relaxation [23] or by the k-shortest paths algorithm [2]. To overcome the limitation imposed by the discrete grid, a purely continuous state space is used in [6]. However, such an accurate description of the complex task leads to a highly non-convex optimi-

sation problem which is minimised locally by gradient descent augmented with heuristic discontinuous jumps. A more elegant discrete-continuous energy was later proposed in [24], where both trajectory estimation and data association are handled simultaneously by minimising a single objective.

The main motivation for designing such complex objective functions [6, 19, 21, 24] is to describe the problem at hand as accurately as possible. Although they obtain state of the art results, they are difficult to optimise and often become trapped in local minima. In practice, this manifests in tracking errors such as fragmented trajectories or confused target identities. In this work we focus on overcoming these errors by applying recent results from single target tracking.

Model-free tracking.

Recent advances in visual single object tracking [25–27] have also adopted the tracking by detection paradigm. However, rather than train the detector offline, so-called *model-free* trackers train a classifier to separate the target from its background, using positive and negative training examples gathered while tracking. This has the advantage of requiring only initialisation in a single frame and of training a detector specifically for the current appearance of the target. Several methods have been applied to the task, including multiple instance learning [28], structured output learning [26], metric learning [29] and kernel methods [27].

In general, model-free tracking methods are successful over short time periods but their performance degrades over longer time spans, or when target appearance changes significantly. By using them to correct short term errors in long term tracks obtained by global methods, we play to the strength of these two different approaches. Because the model-free tracker operates only on the output of the global tracker, it is independent of its implementation and can therefore be combined with any of the above tracking frameworks. The final result is still obtained by optimising the global objective function; the short term tracks are simply used to generate plausible hypotheses, which the optimiser can use to break out of local minima.

Other recent work has also demonstrated the use of single target visual trackers within multi-target tracking. In [10], contours of multiple objects of arbitrary shape are represented using level-sets and an underlying generative model determines location, depth ordering and segmentation of each target. Similarly, a level-set tracker is also applied in the context of pedestrian tracking from a moving camera in [11], where sparse person detections are augmented with the temporally varying target contours provided by the low-level tracker. Izadinia et al. [30] detect pedestrians using the deformable part-based model [31] and in addition to tracking entire people, trajectories of their individual body parts are recovered. In [12], multi-target tracking is based on both, detections from an offline object detector and a visual tracker output. The decision on which cue to use is made based on a pre-trained model using several

features such as detector response or optic flow. Zhang et al. [32] also propose to couple several individual trackers by enforcing to preserve the spatial structure between all targets over time. While this may help to resolve data association in certain cases where objects tend to exhibit similar motion patterns, it is not generally applicable to arbitrary people tracking, in particular sports videos with abrupt and erratic target motion.

Our method is different from previous work in the following aspects: (i) We exploit the power of visual trackers explicitly in difficult situations. To this end, we localise difficult situations in the space-time volume and use the output of multiple coupled single target trackers to generate a strong set of local hypotheses. (ii) We present a local data association scheme for single target trackers. To avoid clumping and identity switching between individual trackers, we follow a simple, yet effective technique based on bipartite graph matching. (iii) We integrate the output of single target trackers into a global energy minimisation method. To avoid potential drift caused by online learned trackers, the local solution is verified in the global context using a robust multiple target objective.

10.3 Multi-target Tracking by Energy Minimisation

In this work, we follow the recent trend and address multi-target tracking by minimising a highly complex energy function. We use the discrete-continuous formulation proposed in [24]. Note, however, that our method is generic and does not rely on any specific formulation of the underlying objective function.

A weakness of any non-convex global objective is that it may become trapped in local minima, which results in fragmented or incorrectly associated tracks. To remedy this, we propose to focus explicitly on those solution regions that are most likely to be erroneous and to guide the optimisation toward alternative solutions using single target tracking with local data association. The entire algorithm is summarised in Algorithm 1, and the individual steps are illustrated in Figure 10.1.

We now describe in more detail each of the steps in the algorithm.

10.3.1 Global data association

In our formulation, multi-target tracking is performed by optimising a discrete-continuous objective, where both data association and trajectory estimation are solved for by minimising a single energy function. Given a set of target detections $\mathbf{D} = d_1, \ldots, d_D$ within a video sequence of F frames, the goal is to find the most likely solution for assigning a unique target identifier to each detection, and at the same time to estimate a continuous trajectory for each target.

Algorithm 1: Tracking multiple targets by local and global optimisation

input : Initial global solution S (Sec. 10.3.1)
output: Final trajectories
while $\neg$ *converged* **do**
  Find next error Ξ in current solution S (Section 10.3.2)
  Optimise locally within spatio-temporal neighbourhood of Ξ to obtain $\hat{S}$ (Sec. 10.3.3, 10.3.4)
  Stitch partial solution $\hat{S}$ into global solution S, if global cost is reduced (Sec. 10.3.5)
end

Following the notation of [24], we represent the state space by two sets of variables. A discrete set $\mathbf{f} = f_i, \ldots, f_D$ determines the data association, where each variable takes on a label l from the label set $\mathbf{L} = \{1, ..., L, \varnothing\}$ which corresponds to a specific target (or a false alarm). A set of continuous variables $\mathcal{T}$ describes the shape of all trajectories under consideration, where each trajectory is represented by piecewise polynomials.

The discrete part of the energy is posed as a graphical model with unary and pairwise potentials and label costs [33]:

$$E(\mathbf{f}) = \phi_d + \psi_{d,d'} + h_{\mathbf{f}} + h^{\mathsf{X}}_{\mathbf{f}}, \tag{10.1}$$

while the continuous part controls the trajectories:

$$E(\mathcal{T}) = \phi_{\mathcal{T}} + h_{\mathbf{f}} + h^{\mathsf{X}}_{\mathbf{f}}. \tag{10.2}$$

In a nutshell, the unary (or data) terms ϕ measure how well the trajectory hypotheses fit the observations, the pairwise terms ψ enforce spatio-temporal smoothness in the labelling, and the label cost models a prior on individual trajectories ($h_{\mathbf{f}}$), such as target dynamics or track persistence, as well as on pairs of tracks ($h^{\mathsf{X}}_{\mathbf{f}}$) to suppress implausible solutions with strongly overlapping trajectories. The complete energy is then minimised by alternately fixing one set of variables at a time, generating the initial solution S. For more details, we refer the reader to [24].

10.3.2 Error detection

Given an initial solution hypothesis S, our goal is to localise errors within this solution and correct them. Several types of local error may exist, including split tracks, swapped identities and merged trajectories. The importance of each error type is application specific, but to demonstrate our approach, we focus only on the most obvious error type that is also convenient to detect, namely an interrupted trajectory. Under the assumption that the scene does not contain any doors or large scene occluders where people may disappear indefinitely, a

target that enters the field of view must ideally remain tracked until it leaves the scene. Therefore, any trajectory $\mathcal{T}_i$ that terminates prematurely and not close to the image border is considered a candidate for improvement. In practice, this is likely to overestimate the number of locations at which errors may occur. This does not detract from the final solution as, in the case of a genuine track endpoint, all hypothesised track joins are likely to result in a higher overall cost, and therefore the initial solution will be unchanged.

Let $\mathbf{x}_i^t$ be the (x, y) location of target i in frame t. Further, let t_i^* denote either the first or the last frame in which target i exists. An error $\Xi = \{x_i, y_i, t_i^*\}$ is possibly present at the spatio-temporal location $\mathbf{x}_i^{t_i^*}$ if and only if $1 < t < F$ *and* $\beta(\mathbf{x}_i^{t_i^*}) > \tau$. $\beta(\cdot)$ computes the distance to the closest image border and τ is a margin where trajectories are allowed to terminate, which is set to 100 pixels in our experiments.

10.3.3 Choosing the local optimisation region

To optimise the solution locally, we consider a spatio-temporal window around each detected error. In particular, we optimise over the temporal window $\mathbf{\Omega} = \{t - k, \ldots, t + k\}$, where k is fixed to 10 frames in our experiments. This time span is usually long enough to resolve an ambiguity, but still local enough to rely on the output of a visual tracker.

It remains to determine which existing trajectories are involved in the current error and should be re-estimated within $\mathbf{\Omega}$. On one hand, it is desirable to reduce the current problem to the smallest possible subset to enable efficient optimisation. On the other hand, discarding too many concurrent trajectories may lead to conflicts in the later step when the two solutions are to be merged. In a typical setting, trajectories that are far apart from one another are independent. A reasonable trade off therefore is to only consider a small subset of trajectories $\mathcal{T}^* \subset \mathcal{T}$ which is within a neighbourhood $\mathbf{\Sigma}$ of the error Ξ. To determine the neighbourhood, we create a short auxiliary trajectory $\hat{\mathcal{T}}$ by tracking the target back and forth within the temporal window $\mathbf{\Omega}$, initialised from the error Ξ. To reduce the state space for the optimisation while at the same time not ignoring important dependencies, we consider only those detections d_i that are within a certain radius of $\hat{\mathcal{T}}$ during the local optimisation (*cf.* Fig. 10.1 (c)). Formally, the set of target candidates is reduced to

$$\hat{\mathbf{D}} = \{d_i | t_i \in \mathbf{\Omega}, \|\mathbf{d}_i - \hat{\mathcal{T}}^{t_i}\| < 2s\}, \tag{10.3}$$

where $\mathbf{d}_i$ denotes the spatial and t_i the temporal location of detection i, respectively, and s is the target size.

10.3.4 Local optimisation

In principle, any existing method can be used to find a plausible solution within the spatio-temporal neighbourhood of the detected error. To guide the optimisation into more promising regions, we exploit single target visual trackers in

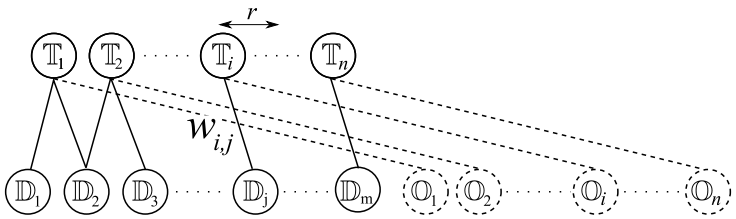


Fig. 10.2: Example of the bipartite graph that must be solved for each frame. Each tracker $\mathbb{T}_i$ is connected to all detections $\mathbb{D}_j$ that lie within its search radius and to one occlusion node $\mathbb{O}_i$.

combination with local data association to generate likely trajectory hypotheses. To this end, we initiate a tracker $\mathbb{T}_i$ from each terminating point of each trajectory $\mathcal{T}_i$ in $\mathcal{T}^*$ that is involved in the error Ξ. In our experiments we employ a recent tracker by Henriques et al. [27]. We use the implementation distributed by the authors. In practice, its high robustness and speed make it feasible to quickly generate many short term track hypotheses, although again other single target trackers could also be used. The resulting tracks form a set of strong candidates for selection in the optimisation procedure.

Traditional single target tracking-by-detection algorithms consider only the single detection in each frame with maximum classification score. In order to solve ambiguous situations where several trackers may detect the same target, we extend this approach by including other possible targets. We include all non-overlapping detections whose classification score is more than 10% of the maximum classification score for the individual tracker. The task of local data association is then to optimise the associations between a set of individually trained trackers $\mathbb{T}$ and the set of detections $\mathbb{D}$ for each frame. By representing this association problem in a bipartite graph we are able to find an optimal solution using the Hungarian algorithm. In order to handle occlusions we also introduce an occlusion node for each tracker, which accounts for fully occluded targets. Figure 10.2 illustrates an example of the bipartite graph for one frame.

Each tracker $\mathbb{T}_i$, $i = 1, \ldots, n$, is initialised, and linked to m detections $\mathbb{D}_1, \ldots, \mathbb{D}_m$ and its respective occlusion node $\mathbb{O}_i$. Edges between trackers and detections with a distance larger than the search radius r have weights zero and are therefore omitted in Figure 10.2. The size of r is chosen as the mean of the height and width of the target. The weight assigned to each edge combines the appearance measure given by the classification score and a proximity measure that penalises large spatial jumps between consecutive frames:

$$w_{i,j} = s_{i,j} \cdot p_{i,j}, \tag{10.4}$$

where $s_{i,j}$ is the classification score for tracker i evaluated on target j, scaled to $[0, 1]$. $p_{i,j}$ is a linear proximity measure between the last detection of tracker

i and target j and is defined as

$$p_{i,j} = \frac{1}{r} \max\left(0, r - \|\mathbb{T}_i^{t-1} - \mathbb{D}_j\|\right). \tag{10.5}$$

The proximity measure is used as a simple random walk motion model. Particularly in sports the motion may be abrupt, therefore, we choose this zero displacement model rather than assuming constant velocity.

Edges connecting a tracker to its occlusion node are assigned a low weight, which is empirically chosen as 8% of the maximum classification score in order to be lower than real detections.

10.3.5 Combining local and global solutions

To stay consistent with the overall formulation, we minimise the same discrete-continuous objective function as is used to evaluate the quality of the complete solution on the spatio-temporal subset $\{\mathbf{\Sigma}, \mathbf{\Omega}\}$ using track hypotheses from Section 10.3.4. After the optimisation, the resulting solution $\hat{S}$ replaces the original solution within $\{\mathbf{\Sigma}, \mathbf{\Omega}\}$ if the overall energy $E(\hat{S} \cup \tilde{S})$ is decreased. $\tilde{S}$ is obtained by simply removing all partial trajectories from S that lie within the spatio-temporal neighbourhood $\{\mathbf{\Sigma}, \mathbf{\Omega}\}$ of the error.

10.4 Experiments

Datasets.

We demonstrate our approach on eight different sequences. The first set consists of six publicly available videos including the PETS 2009 benchmark [34][1] and TUD Stadtmitte [35]. All videos show pedestrians in a single view but they exhibit a large variation in person count, camera viewpoint and motion patterns. Since the camera calibration is available for this dataset, we perform tracking on the ground plane in world coordinates.

As well as evaluating on standard benchmarks, we also demonstrate the performance on difficult sports tracking data. In particular we show tracking on two sequences in the challenging sport of Australian Rules Football (AFL), in which there is regular and frequent crowding of players and contact between them, making it a very difficult tracking problem. We make this new dataset including the ground truth annotations and the detections used in this work publicly available[2].

Metrics.

Quantifying performance of multiple target tracking is a notoriously difficult task [36]. Ambiguities in annotations, assignments strategies and metric de-

[1] Sequences: S2L1, S2L2, S2L3, S1L1-2, S1L2-1

[2] http://research.milanton.net/data

scriptions prohibit a purely objective evaluation. Here we follow the most widely used strategy and report several metrics for all our experiments. Next to standard precision and recall figures we report the CLEAR MOT metrics [37], which consists of tracking accuracy (MOTA) and tracking precision (MOTP). The former combines three error types: false positives, missed targets and identity switches, into a single number such that zero errors corresponds to 100%. The latter measures the localisation error of the tracker w.r.t. the annotated ground truth. Moreover, we also show the number of correctly recovered trajectories as proposed in [38]. A target is considered mostly tracked (MT), if it is correctly detected in over 80% of frames within its time span. Similarly, a mostly lost (ML) trajectory is only recover in 20% of frames or less. Finally, the numbers of track fragmentations and identity switches are stated for completeness.

Before presenting the overall tracking performance of our system, we discuss the importance of local data association for single target trackers and illustrate the potential of our locally driven optimisation scheme measured by the reduction of the total cost. We then provide an extensive quantitative evaluation on various challenging sequences and compare our results to several state-of-the-art methods.

10.4.1 Local data association

Let us first qualitatively demonstrate the effect of local data association using multiple model free trackers in situations with a high presence of occlusions. Figure 10.3 shows two comparisons between tracking with and without local data association. The first sequence is from the PETS 2009 S2L2 dataset, and the second one is a challenging situation from an AFL game. The images are cropped for better visibility. In all cases model-free trackers are initialised for each target depicted with bounding boxes in the left most image. The 1$^{\text{st}}$ and 3$^{\text{rd}}$ rows show the results of running the two, respectively three trackers individually, without local data association. In rows two and four the results are obtained by including the Hungarian data association described in Section 10.3.4. The results show clearly that the individual trackers are prone to drift in settings with multiple persons. In row 1 the identity switches and the blue target is lost after occlusion. By including local data association these situations are resolved, and the two targets are correctly tracked even after full occlusions. In the 3$^{\text{rd}}$ row without data association the three trackers clump together and follow the same person which yields the highest classifier score for each of the trackers after the occlusion. The local data association shown in the 4$^{\text{th}}$ row again resolves this situation and keeps tracking the three individual persons while maintaining their correct identities.

To quantify the importance of multiple tracker reasoning, we minimise the global objective with and without explicit local data association (no LDA) for proposal generation. Experimental results are reported in the following sections.

Fig. 10.3: Visual trackers without (1st and 3rd rows) and with (2nd and 4th rows) local data association. See text for details.

10.4.2 Energy minimisation

To verify the potential of our approach, we compare the magnitude of the initial solution S of the overall energy to the final solution obtained after including local tracks. Figure 10.4 shows the relative energy decrease for various sequences. The energy is scaled in each case such that the initial point, which is obtained by [24], corresponds to 100%. It is important to note that we minimise the exact same energy without introducing new detections. By focusing on erroneous regions and by exploiting model-free trackers for better hypothesis generation, our proposed local optimisation can find a lower global cost in nearly every iteration and an overall reduction of over 50% in some cases. Dotted lines for the AFL sequences show the energy reduction using proposals of independent single target trackers without local data association. One iteration takes approximately one second to compute on a standard PC. We set the maximum number of iteration to 150 in all our experiments.

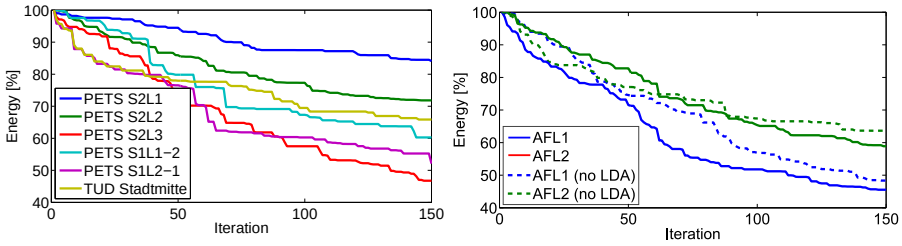


Fig. 10.4: Minimising a global energy function by focusing on local optimisation windows. The dotted plots on the right hand side depict the energy minimised by our scheme by only using independent single target trackers *without* local data association (no LDA).

Table 10.1: Quantitative results on two AFL sequences. Best result across all methods is highlighted in bold face for each measure.

Method	MOTA	MOTP	MT	ML	Frag.	ID sw.	Precision	Recall
FFP Detector [39]	–	–	–	–	–	–	65.4%	55.0%
SMOT [40]	16.7%	60.8%	2	3	**38**	**14**	59.8%	52.0%
DCO [24]	29.7%	63.3%	3	**2**	93	97	70.9%	56.3%
ours (no init)	32.0%	**64.1%**	6	**2**	54	54	67.4%	64.5%
ours (no LDA)	39.0%	63.6%	6	**2**	44	27	72.1%	64.2%
ours (full)	**41.4%**	63.6%	**7**	**2**	39	22	**73.2%**	**65.8%**

10.4.3 Quantitative evaluation

AFL sports data.

We first demonstrate quantitative performance of our approach on the two sports sequences. To obtain candidate detections, we trained a person detector based on fast feature pyramids [39] using only one single image as training data resulting in moderate precision and recall. Table 10.1 shows the detector's performance as well as tracking results from two recent multi-target tracking methods. The similar-appearance multiple object tracker (SMOT) [40] is specifically designed to address situations shown in these sports sequences with similar target appearance by relying only on motion similarity and using a generalised linear assignment to reconstruct long-term tracks. While this method shows excellent performance with no or little detection noise, it struggles to correctly infer plausible trajectories in a realistic challenging setting. The second baseline is a recent energy minimisation-based method (DCO) [24], which can eliminate many false positive detections. However, due to the complex formulation of its objective, the optimisation reaches only a moderate local minimum with many short tracks leading to a high number of interrupted trajectories and identity switches.

The second part shows three variants of our proposed method. The first one (no init) is our optimisation strategy starting from the trivial solution, where each detection is considered an error (or equivalently a single-frame track). Note that we are able to outperform other methods by appying our customized optimisation scheme. The second strategy (no LDA) uses [24] as initialisation

Table 10.2: Comparison to previous methods on a standard benchmark (PETS, TUD). The results are averaged over six sequences.

Method	MOTA	MOTP	MT	ML	Frag.	ID sw.	Precision	Recall
HOG/HOF Det. [41, 42]	–	–	–	–	–	–	79.5%	62.2%
DP [16]	46.0%	**64.7%**	8	11	165	204	91.7%	55.8%
KSP [2]	41.7%	62.8%	8	20	**10**	**18**	91.6%	46.8%
DCO [24]	55.7%	63.6%	11	**9**	49	43	93.0%	61.6%
ours	**56.9%**	64.1%	**13**	10	40	48	**93.4%**	**62.8%**

but does not involve local data association for hypothesis generation as described in Section 10.3. Finally, by applying our full method using localised optimisation with visual trackers, we are able to further minimise the objective function, which is also reflected in the superior tracking performance.

Public benchmark.

Our second set of experiments involves a public tracking benchmark. Table 10.2 shows a quantitative comparison of our proposed strategy to previous methods: A network flow-based approach solved with dynamic programming (DP) [16], globally optimal tracking on a discrete grid (KSP) [2] and the same energy minimisation formulation as before [24]. All numbers are computed using code provided by the authors, publicly available detections, ground truth and evaluation scripts[3]. Note that the slightly higher absolute number of ID switches is a result of incorrectly bridging or extending interrupted trajectories. However, the positive effect of recovering more tracks (MT) and thereby increasing the recall outweighs, yielding higher overall accuracy.

Although we outperform state-of-the-art methods on this benchmark, the improvement is less prominent than in the AFL case. One reason for this behaviour may be that the detection quality is poorer on the sports sequences due to large deformations and small target size (*cf.* Tab. 10.1 and Fig. 10.5), yielding a more complex optimisation problem with more local minima. It is also possible that [24] finds a solution much closer to the global optimum on the public sequences, which may indicate that it is well suited to the benchmark but shows limitations on novel data.

10.4.4 Qualitative results

Finally, Figure 10.5 illustrates qualitative results on three sequences. Each row shows three frames from AFL1, AFL2, and PETS S2L2, respectively. Note that our method is able to correctly identify nearly all targets even in extremely challenging conditions with substantial levels of multiple occlusions. Also note how the potential of using visual model-free trackers within a traditional multi-person tracking setting is unfolded in situations with extensive pose variation, such as demonstrated by the cyan (ID 33) and the blue (ID 20) targets in the

[3] Note that the corrected numbers are reported for [24], which differ from the original publication.

Fig. 10.5: Exemplar frames from two AFL clips and the PETS S2L2 sequence. Note that using model-free trackers allows one to maintain the identity of a player even during severe deformations and pose changes (*cf.* the blue target (ID 20) in the second row).

first and second row, respectively. Please refer to the supplemental video for further visual results.

10.5 Conclusion

We proposed a simple yet effective method to optimise highly complex objectives for multiple target tracking by focusing explicitly on correcting errors locally. A local data association technique combined with a set of visual object trackers is able to drive the optimisation into much better minima reducing the energy by over 50% and consequently leading to superior solutions. We demonstrate the validity of our approach on particularly challenging sports sequences and public benchmark data achieving state of the art performance.

In future work we plan to more thoroughly investigate different error types and their influence on the final solution. It may also be possible to design even more accurate and more complex objective functions that better approximate the true state but still remain tractable using our local optimisation strategy.

Acknowledgements

We gratefully acknowledge the financial support of the Australian Research Council through Laureate Fellowship FL130100102 to IDR.

References

[1] L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," in *CVPR 2008*.

[2] J. Berclaz, F. Fleuret, E. Türetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *IEEE T. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1806–1819, Sep. 2011.

[3] B. Yang and R. Nevatia, "An online learned CRF model for multi-target tracking," in *CVPR 2012*, pp. 2034–2041.

[4] A. R. Zamir, A. Dehghan, and M. Shah, "GMCP-Tracker: Global multi-object tracking using generalized minimum clique graphs," in *ECCV 2012*, vol. 2, pp. 343–356.

[5] J. Liu, P. Carr, R. T. Collins, and Y. Liu, "Tracking sports players with context-conditioned motion models," in *CVPR 2013*, pp. 1830–1837.

[6] A. Milan, S. Roth, and K. Schindler, "Continuous energy minimization for multitarget tracking," *IEEE T. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 58–72, 2014.

[7] J. Vermaak, A. Doucet, and P. Pérez, "Maintaining multi-modality through mixture tracking," in *ICCV 2003*, pp. 1110–1116.

[8] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Robust tracking-by-detection using a detector confidence particle filter," in *ICCV 2009*.

[9] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *CVPR 2013*, pp. 2411–2418.

[10] C. Bibby and I. Reid, "Real-time tracking of multiple occluding objects using level sets," in *CVPR 2010*, pp. 1307–1314.

[11] D. Mitzel, E. Horbert, A. Ess, and B. Leibe, "Multi-person tracking with sparse detection and continuous segmentation," in *ECCV 2010*, vol. 1, pp. 397–410.

[12] X. Yan, X. Wu, I. A. Kakadiaris, and S. K. Shah, "To track or to detect? An ensemble framework for optimal selection," in *ECCV 2012*, vol. 5, pp. 594–607.

[13] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[14] S. J. Julier and J. K. Uhlmann, "A new extension of the kalman filter to nonlinear systems," in *International Symposium on Aerospace andDefense Sensing, Simulation and Controls*, 1997, pp. 182–193.

[15] H. Jiang, S. Fels, and J. J. Little, "A linear programming approach for multiple object tracking," in *CVPR 2007*.

[16] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *CVPR 2011*.

[17] J. a. Henriques, R. Caseiro, and J. Batista, "Globally optimal solution to multi-object tracking with merged measurements," in *ICCV 2011*.

[18] A. A. Butt and R. T. Collins, "Multi-target tracking by lagrangian relaxation to min-cost network flow," in *CVPR 2013*.

[19] B. Leibe, K. Schindler, and L. Van Gool, "Coupled detection and trajectory estimation for multi-object tracking," in *ICCV 2007*.

[20] Z. Wu, A. Thangali, S. Sclaroff, and M. Betke, "Coupling detection and data association for multiple object tracking," in *CVPR 2012*.

[21] W. Brendel, M. R. Amer, and S. Todorovic, "Multiobject tracking as maximum weight independent set," in *CVPR 2011*.

[22] Z. Wu, T. H. Kunz, and M. Betke, "Efficient track linking methods for track graphs using network-flow and set-cover techniques," in *CVPR 2011*.

[23] J. Berclaz, F. Fleuret, and P. Fua, "Multiple object tracking using flow linear programming," in *12th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (Winter-PETS)*, Dec. 2009.

[24] A. Milan, K. Schindler, and S. Roth, "Detection- and trajectory-level exclusion in multiple object tracking," in *CVPR 2013*.

[25] Z. Kalal, J. Matas, and K. Mikolajczyk, "P-N Learning: Bootstrapping binary classifiers from unlabeled data by structural constraint," in *CVPR 2010*.

[26] S. Hare, A. Saffari, and P. H. S. Torr, "Struck: Structured output tracking with kernels," in *ICCV 2011*, pp. 263–270.

[27] J. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *ECCV 2012*, vol. 4, pp. 702–715.

[28] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in *CVPR 2009*.

[29] X. Li, C. Shen, Q. Shi, A. Dick, and A. v. d. Hengel, "Non-sparse linear representations for visual tracking with online reservoir metric learning," in *CVPR 2012*, pp. 1760–1767.

[30] H. Izadinia, I. Saleemi, W. Li, and M. Shah, "$(MP)^2T$: Multiple people multiple parts tracker," in *ECCV 2012*, vol. 6, 2012, pp. 100–114.

[31] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE T. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, 2010.

[32] L. Zhang and L. van der Maaten, "Structure preserving object tracking," in *CVPR 2013*, pp. 1838–1845.

[33] A. Delong, A. Osokin, H. N. Isack, and Y. Boykov, "Fast approximate energy minimization with label costs," *Int. J. Comput. Vision*, vol. 96, no. 1, pp. 1–27, Jan. 2012.

[34] J. Ferryman and A. Shahrokni, "PETS2009: Dataset and challenge," in *11th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, Dec. 2009.

[35] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3D pose estimation and tracking by detection," in *CVPR 2010*.

[36] A. Milan, K. Schindler, and S. Roth, "Challenges of ground truth evaluation of multi-target tracking," in *2013 IEEE CVPR Workshops (CVPRW)*, Jun. 2013, pp. 735–742.

[37] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *Image and Video Processing*, vol. 2008, no. 1, pp. 1–10, May 2008.

[38] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: Hybridboosted multi-target tracker for crowded scene," in *CVPR 2009*.

[39] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE T. Pattern Anal. Mach. Intell.*, 2014, to appear.

[40] C. Dicle, M. Sznaier, and O. Camps, "The way they move: Tracking multiple targets with similar appearance," in *ICCV 2013*.

[41] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR 2005*, pp. 886–893.

[42] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *CVPR 2010*.

Part V

Smart City applications

Until this part we have been focusing on the analysis of sports facilities and sports players. This part will open up for a more broad use of the technology and present the results from analysing humans in the city.

In the Smart City, digital technology monitors and regulates the installations of a city, for a more efficient, healthy, and liveable environment. With cameras and automatic computer vision software, it is possible to analyse large amounts of data in a non-intrusive manner. The privacy conserving nature of the thermal sensor makes it well suited for use in public installations. In the first chapter of this part we show how thermal imaging performs in five different applications. The following two chapters go into more detail on two different applications, using the same basic tracking framework.

The first chapter of this part consists of a paper accepted for journal publication and the two following chapters were originally published in conference proceedings:

Rikke Gade, Thomas B. Moeslund, Søren Zebitz Nielsen, Hans Skov-Petersen, Hans Jørgen Andersen, Kent Basselbjerg, Hans Thorhauge Dam, Ole B. Jensen, Anders Jørgensen, Harry Lahrmann, Tanja Kidholm Osmann Madsen, Esben Skouboe Bala and Bo Ø. Povey, "Thermal Imaging Systems for Real-Time Applications in Smart Cities," *International Journal of Computer Applications in Technology*, accepted for publication.

Esben Skouboe Poulsen, Hans Jørgen Andersen, Ole B. Jensen, Rikke Gade, Tobias Thyrrestrup and Thomas B. Moeslund, "Controlling Urban Lighting by Human Motion Patterns - results from a full scale experiment," *Proceedings of the ACM International Conference on Multimedia*, pp. 339–347, October 2012.

Søren Zebitz Nielsen, Rikke Gade, Thomas B. Moeslund and Hans Skov-Petersen, "Taking the temperature of Pedestrian Movement in Public Spaces," *Transportation Research Procedia - Conference on Pedestrian and Evacuation Dynamics*, vol. 2, pp. 660–668, October 2014.

Chapter 11

Thermal Imaging Systems for Real-Time Applications in Smart Cities

Rikke Gade, Thomas B. Moeslund, Søren Zebitz Nielsen, Hans Skov-Petersen, Hans Jørgen Andersen, Kent Basselbjerg, Hans Thorhauge Dam, Ole B. Jensen, Anders Jørgensen, Harry Lahrmann, Tanja Kidholm Osmann Madsen, Esben Skouboe Bala and Bo Ø. Povey

The paper is accepted for publication in
International Journal of Computer Applications in Technology

© 2014 Inderscience Enterprises Ltd.
The layout has been revised.

Abstract

In the modern world, cities need to keep up with the demand for mobility, efficient infrastructure and environmental sustainability. The future Smart Cities use intelligent Information and Communication Technologies to raise the quality of life. This includes computer vision as one of the main technologies. It can observe and analyse human activities from a distance in a non-invasive manner. Traditional computer vision utilises RGB cameras, but problems with this sensor include its light dependency, and the privacy issues that can be raised by people being observed. In this paper, we propose the use of thermal imaging in real-time Smart City applications. Thermal cameras operate independently of light and measure the radiated infrared waves representing the temperature of the scene. In order to showcase the possibilities, we present five different applications which use thermal imaging only. These include both indoor and outdoor scenarios with the purposes of people detection, counting and tracking, as well as one application for traffic safety evaluation.

11.1 Introduction

Contemporary urbanisation on a global scale has led to a number of challenges in terms of environmental stress, continuing migration, demographic shifts and urban liveability. In the literature on the application of digital technologies to these vast urban challenges, the notion of a 'Smart City' is frequently seen [1, 2]. If one set aside the policy and branding value of an overarching, but rather simplifying term, the notion of a 'Smart City' carries the potential for engaging with these urban challenges. The important feature to notice is that the contemporary city no longer can be comprehended as a bounded and mononuclear unit; rather, cities are connected (or dis-connected) in globally reaching networks that create new dynamics of mobility, regardless of whether these are of the nature of moving goods, ecological flows or in-migrating people [3]. But next to these vast new material transformations, the 'network city' has also become over-layered with new types of digital technology, leading some scholars to speak of a 'sentient city' [4] or a 'programmable city' [5]. The essence of this development is that networked technologies are now able to track and detect different types of movement (container, information, cars, or people) to such an extent that we may speak of a new 'digital layer' to the city [3]. Obviously, many scenarios may play out from this. Some would be focusing on the issue of surveillance and control, whereas others would put focus on the potential for these new technologies to optimise the flow of people, goods and information, as well as to engage citizens more in their cities and to create better environmental solutions. However, in order to explore the potential of these new networked technologies, and thus apply them in 'smart' ways, basic research needs to be undertaken. One must explore what the new networked technologies may afford in solutions to the grand challenges facing the contemporary cities globally. As

we are beginning to see cities where real-time location aware data is being created on a continuous basis, we may start to ask how such data can inform better systems of traffic planning, city governance, energy provision and other types of urban utilities [6]. But moreover, the new 'digital layer' opens up a discussion of whether it is possible to involve citizens more in the social life of public spaces, if the new technologies hold the potential to create new aesthetic experiences 'on top' of the existing ones; and if real-time data fed back to city inhabitants will create a whole new experience of the city. These issues are relevant dimensions of a new research agenda on the 'Smart City', and in this paper we will be focusing on one particular combination of technologies that we think need to be researched for their undiscovered potentials.

Computer vision technologies have great potential in smart city applications, due to the non-intrusive nature of the sensor. With cameras, it is possible to record data at a distance and at real-time. There are, however, some problems with regular RGB cameras. First of all, the deterrent effects of surveillance and control caused by cameras are high, as it is possible to visually recognise and follow people[1]. Moreover, robust systems can be very hard to develop, as cameras capture reflected lighting. Thus, they depend on sufficient light in order to capture anything, and the visual perception of things changes with the lighting. In the context of smart cities, these problems are critical, as 24 hour operation is assumed in many applications. This is not always possible to achieve with RGB sensors, while the natural lighting changes with the weather and the nights are dark. An alternative sensor, which is still non-intrusive, but also independent of lighting, is the thermal camera; and with this sensor, privacy issues are also eliminated. Thermal cameras capture the long-wavelength infrared radiation, which is radiated from all objects with a temperature above absolute zero. The amount of radiation depicts the temperature of the object, resulting in an image that visualises the temperature of the scene. People and objects with a temperature different from the surroundings can therefore be detected both day and night. Figure 11.1 shows an example of a thermal image from an urban square.

We have in recent years done a large amount of research in thermal imaging for different applications related to the Smart City. The purpose of this paper is firstly, to introduce the thermal sensor and its potential to the computer vision research community. We will compare and discuss the use of thermal and RGB sensors for real-time people detection and tracking. Furthermore, we will showcase five successful use-cases where we have applied computer vision to thermal imagery in the context of Smart Cities.

The following section will discuss thermal technology and compare RGB and thermal cameras. After that, we present five different smart city applications where real-time thermal imaging is applied.

[1] In some surveillance situations, discovery of a person's identity is the ultimate goal; but in general, data acquisition is preferred to be anonymous

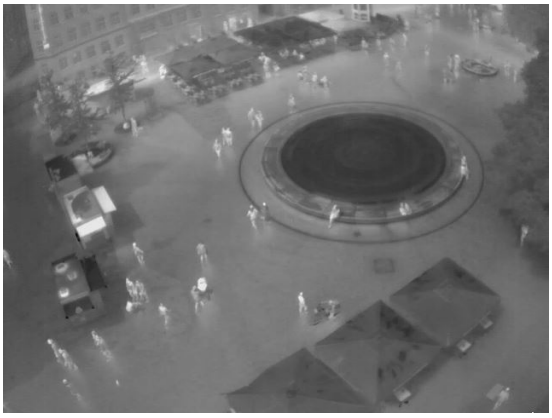


Fig. 11.1: Thermal image overlooking an urban square.

11.2 Thermal Sensors

Originally developed for military purposes, thermal cameras have now reached a reasonable price and have become available for commercial use. The technology has therefore started to be deployed for surveillance purposes, where RGB cameras would typically have been used.

The resolution of thermal sensors is slowly increasing as the technology evolves and new materials are explored. Available today are sensors ranging from cheap 8×8 pixel arrays [7], up to 1280×1024 pixel sensors [8]. The field-of-view ranges from very narrow (approx. 1 degree), to wide angle lenses up to 80 degrees [9]. Some types of cameras have the possibility for optical zoom. The wide angle lenses are very useful in applications like surveillance of indoor rooms and urban spaces. However, the selection of wide angle lenses are small and the price is very high, due to the expensive lens material germanium. With very narrow field-of-view lenses, it is possible to detect people several kilometres away. This is very useful, especially for border and coastal surveillance.

For commercial use, the best known types of thermal cameras are the hand-held cameras for building inspections. However, thermal surveillance cameras are becoming very popular, due to their independence of lighting. These cameras are very similar to regular RGB surveillance cameras in terms of size, look and interfaces [10]. RGB cameras are still significantly cheaper than thermal cameras, though.

Thermal network cameras with data communication over IP can be part of larger camera networks, and they enable easy data transfer to a computer [11]. Many cameras come with built-in memory or a slot for a memory card, but external computers, and thereby storage, can also be connected to the cameras. Some cameras have small built-in processing units, where simple image processing algorithms can be programmed; e.g., motion detection and cross

line detection. Real-time online processing of a video with more demanding algorithms is possible when a computer is connected.

11.2.1 Segmentation of people

Our main purpose of using computer vision in Smart Cities is to automatically, and in real-time, detect and track the movements of humans [12, 13]. Therefore, we will focus on the possibility of people tracking and detection in each of the image modalities and start with a general overview of detection methods.

In almost any application of computer vision figure-ground segmentation is needed to locate the desired object(s) in the images. Mainly, two approaches are widely used for detecting humans: pixel-based detection and object-based detection [14]. Pixel-based detection methods consider each pixel individually, e.g. by comparing to a background model. The basic idea is to compare each pixel to a background model, and if the difference exceeds a given threshold, the pixel is classified as foreground. However, especially in outdoor scenes, obtaining a valid background model is challenged by the shifting sun, clouds, moving branches of trees, etc. Different ways of modelling a changing background, as well as updating it appropriately, is still being researched. In applications with a moving camera, it might be impossible to model the background, though.

Image thresholding is applied after background subtraction, but can also be applied to original images of different modalities. Depending on the application, thresholding methods vary from the simplest constant threshold value over dynamic and automatic methods [15] to bio-inspired algorithms [16].

Object-based methods detect either entire objects or major parts in case of a part-based model. Often these methods run by translating a window over the image and calculate the likelihood of each window containing a human [14]. The Histogram of Oriented Gradients (HOG) detector [17] is one example of a very popular object-based detector, searching for a learned object shape.

The pixel-based methods detects everything that is or has been moving. This approach is fast, but it will often require post processing to filter noise and unwanted objects from the detections.

Object-based methods on the other hand are designed to detect a specific type of object and normally needs no post processing. However, the output is usually bounding box locations, compared to the more precise silhouettes produced by pixel-based methods. The detection rates for HOG degrade rapidly when the resolution decreases (people less than 80 pixels tall), or if people are partly occluded [18]. Furthermore, object-based methods are generally computationally expensive, and often require GPU-based implementations in order to perform in real-time.

These limitations of object-based methods affect its use in Smart City applications, since the appearance of people may vary a lot. People are often observed from a far view, resulting in low resolution and different viewing angles of the people. On the other hand, the camera is most often static, reducing the problems of modelling a background for pixel-based methods significantly.

For the applications at hand, with focus on robust real-time performance, pixel-based detection methods seem best suited for detection. In the following section we will test and compare detection in RGB and thermal images.

11.2.2 Comparison of thermal vs. RGB cameras

For any kind of detection, thermal cameras have clear advantages compared to RGB cameras in dark conditions, due to the independence of lighting. Figure 11.2 compares the two modalities from a night scene. Even though parts of the scene are illuminated, the person is very hard to detect in the RGB image. In the thermal image, the person is fully visible.

(a) RGB (b) Thermal

Fig. 11.2: Example of a night scene.

Due to the properties of thermal imaging, detection of people can be easy and fast in the situations where human temperature differs from the surroundings. Here, a thresholding of the images can be sufficient for segmentation. Figure 11.3 shows an example where an automatic threshold method based on Maximum Entropy [15] is applied to a thermal image.

(a) Thermal input (b) Automatic thresholding

Fig. 11.3: Segmentation of people using automatic thresholding.

A similar fast approach which can be applied to both RGB and thermal

images is background subtraction, which we described in section 11.2.1. This approach assumes that only people, or other objects of interest, are moving. Otherwise filtering of the detections must be applied as post processing.

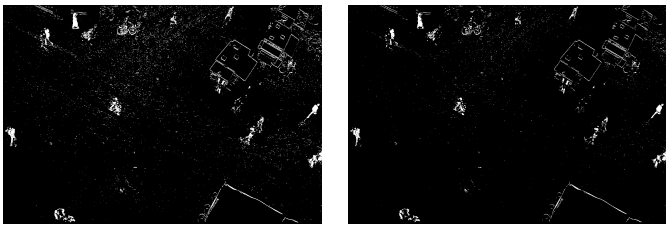
Figures 11.4(a) and 11.4(b) show an example of an urban outdoor space captured simultaneously by an RGB and thermal camera (with slightly different views). Figures 11.5 and 11.6 show the results of background subtraction and binarisation with different threshold values. The threshold values are here chosen manually to illustrate the best possible segmentation and to show the effect of changing the values. In a real application an automatic thresholding method could be applied for choosing the best threshold value in any given frame.

Dark colours of clothes and the relatively small size of people make it very hard to detect people in the RGB image. Furthermore, the ground is partly wet, making it even harder to distinguish people. Using background subtraction, some people can be detected, but as figure 11.5 shows, it is a trade-off between too much noise and missed detections. Using the thermal image, less noise is detected, even with a much lower threshold value of 5. Thus, as seen in figure 11.6, the trade-off is less distinct, and people can more reliably be segmented with background subtraction using a thermal image.

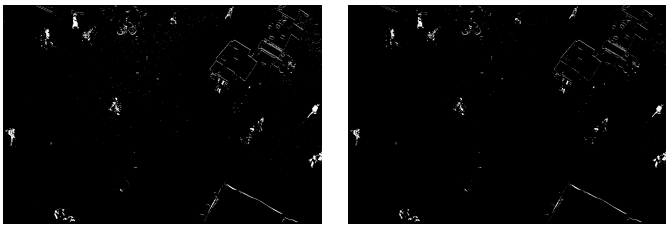
(a) RGB (b) Thermal

Fig. 11.4: An urban scene captured with an RGB and a thermal camera. The cameras have slightly different views. The images are captured on an early summer day with temperatures around 15°C.

Segmentation using background subtraction assumes that it is possible to obtain a reliable background model. Fast changing lighting/temperature conditions can cause problems, and a dynamic background model must be adjusted and updated during run-time [19]. Thermal images have only one channel, compared to three channels of an RGB image. Furthermore, the temperature often changes more slowly than the lighting. This can make it easier to model the background in the thermal domain. Another well-known problem in the RGB domain is the occurrence of shadows. Shadows often cause false detections, as the shadows move just like people. Shadows are related to lighting, and thereby do not exist in the thermal domain. However, thermal radiation



(a) th = 30. (b) th = 40.



(c) th = 50. (d) th = 60.

Fig. 11.5: Background subtraction of the image in figure 11.4(a), then binarisation with threshold values from 30-60. The edges of the carts in the upper right corner are detected due to small vibrations of the camera.

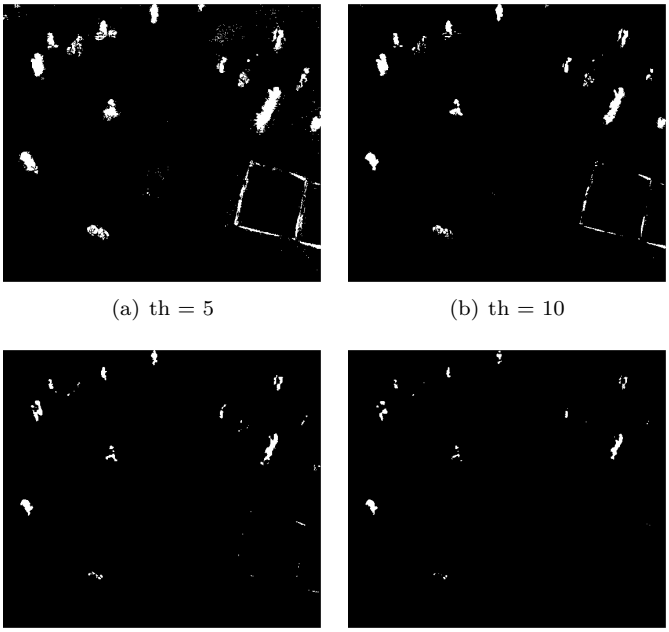
can be reflected in glossy surfaces and cause false detections that are similar to shadows. This is considered a rarer problem in Smart City applications though, as the surfaces are often not reflective.

After detection, some Smart City applications require the object to be tracked. Tracking humans is a complex problem, due to the dynamically changing motion, and often occlusions must be resolved. In these situations, it can be necessary to re-identify people after an occlusion or an ambiguous situation. Recognising individual people is difficult in thermal images, due to the lack of colour and texture information. Thereby, re-identification of people in thermal images is very difficult, making it very hard to solve these ambiguous situations in tracking. Thus, RGB cameras have advantages in complex tracking scenarios.

In table 11.1, the pros and cons of using RGB and thermal cameras for people detection and tracking are summarised.

11.3 Application of Real-time Thermal Imaging

In the remaining part of this paper, we present five different Smart City applications. The purposes are all related to real-time monitoring of the city, in order to optimise mobility, accessibility and safety for the citizens. In all applications, we use one or more uncooled thermal cameras with a resolution



(a) th = 5 (b) th = 10

(c) th = 20 (d) th = 30

Fig. 11.6: Background subtraction of the image in figure 11.4(b), then binarisation with threshold values from 5-30.

of 384 × 288 pixels (AXIS Q1921) or 640 × 480 pixels (AXIS Q1922). Table 11.2 summarises the applications in terms of purpose, equipment, test time, type of scene, technique used and processing time. The test time stated here is the full run time of the set-up, the manually annotated period for quantitative evaluation may be shorter.

11.4 People Counting in Urban Environments

In a Smart City, human movement is automatically registered and analysed. For both real-time and long-term perspectives, this knowledge can be beneficial in relation to urban planning and for shopkeepers in the city. Information in real-time can be used for analysing the current flow and occupancy of the city, while long-term analysis can reveal trends and patterns related to specific days, time or events in the city. In this work, we continuously counted people passing through a pedestrian zone in a city environment during one week. The location is illustrated in figure 11.7, along with a picture of the camera's view.

	Pros	Cons
RGB	Cheap sensors Re-identification possible	Sensitive to light Privacy issues Shadows
Thermal	Easier segmentation Independent of light No privacy issues Single channel images	Re-ident. difficult More expensive Reflections

Table 11.1: Comparison of RGB and thermal cameras for people tracking and detection.

Sec.	Purpose	Camera(s)	Test time	Scene	Technique	Proc. time
11.4	People counting	1 AXIS Q1921 10mm	1 week	Outdoor pedestrians	Image differencing	1.2 ms/f
11.5	People tracking	3 AXIS Q1921 19mm	1 week	Outdoor pedestrians	Background subtraction and Kalman filtering	66 ms/f
11.6	Car and bicycle detection	1 AXIS Q1922 10mm	20 days	Outdoor traffic	Optical flow	41 ms/f
11.7	People counting and activity recognition	3 AXIS Q1922 10mm	1 month	Indoor sports	Background subtraction and Fisher faces	12.5-60 ms/f
11.8	People tracking	1 AXIS Q1922 10mm	2 hours	Outdoor pedestrians	Background subtraction and Kalman filtering	20 ms/f

Table 11.2: Summary of the applications presented in this paper.

11.4.1 Methods

When counting people in a pedestrian zone or a sidewalk, it is assumed that most people are moving across the scene. But being an outdoor scene, it cannot be assumed that people are constantly hotter or colder than the background. The surrounding temperature will change throughout the day, and the sun can heat dark pavement to temperatures hotter than the human body temperature. Therefore in this work, it has been chosen to do segmentation based on image differencing. The activity are then counted solely on the pixel response as opposed to connected silhouettes of people. Double differencing will be used in order to eliminate noise. Figure 11.8 explains the algorithm.

A threshold value of 2 is applied to binarise the image. Figure 11.9 shows an example of an input frame, and the result of double differencing with the two previous frames.

The activity, here represented by white pixels, must be converted into a number of people. From training data, the relation between the amount of activity and the number of people can be calculated. This factor depends on the specific set-up; camera specifications and mounting location. For each new location, a training phase is needed in order to learn this factor. However, it is tested to generalise appropriately over different days and activity levels at the same location. When the velocity of people, and thereby the activity in the frame, is lower, the person will instead stay longer within the region of interest, so that the accumulated activity over a time period will be equal to that of a person moving fast through the scene. The number of people should

Fig. 11.7: Illustration of the location and camera view.

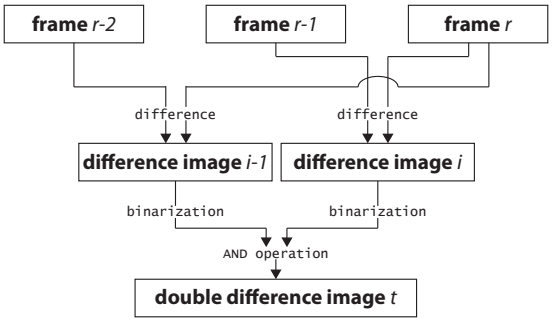


Fig. 11.8: Double differencing algorithm.

therefore be estimated for time windows, rather than momentarily. Choosing a smaller region of interest (ROI) in the image, as illustrated in figure 11.10, will improve the results by reducing the perspective effects in the image. The region is chosen to represent a section of the street as close to the camera as possible. The top-view of the scene in combination with the chosen region of interest implies that occlusions can be neglected.

The ROI is applied by using an AND-operation between a binary image representing the ROI and the current frame.

11.4.2 Results

One full week of video has been captured. Due to the high time consumption of manual video annotation, 13 video sequences of 20 minutes each have been chosen for tests. These video sequences cover different days, times of the day and the level of activity in the video, and they have been manually annotated. Using a leave-one-out cross validation, with twelve video sequences used for training and one for testing in each iteration, the mean accuracy is 90.75 %.

The full week of video has been processed and compared to the results of a commercial system [20]. This system uses observed Bluetooth ID's to estimate the number of people passing a node. This does, however, only represent the

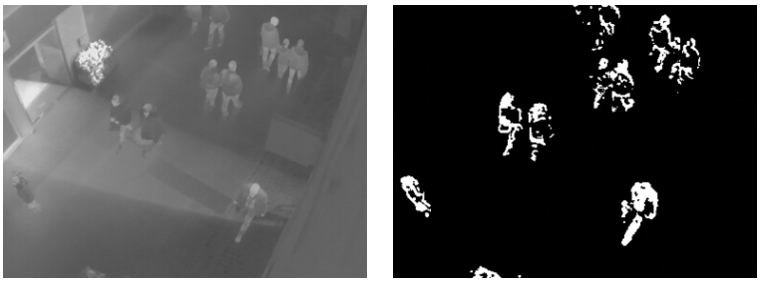


Fig. 11.9: Left: Thermal input image, right: Result of double differencing and thresholding.

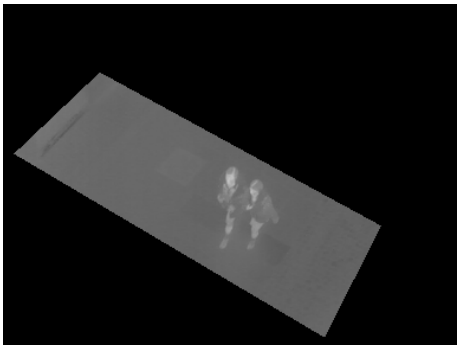


Fig. 11.10: The chosen region of interest.

percentage of people carrying a device with an open Bluetooth connection. We scale the Bluetooth results with a factor of 25.76, in order to compare the results as shown in figure 11.11. The factor of 25.76 has been found experimentally by dividing the mean of our person detections with the mean of detected Bluetooth devices.

None of these results have been manually verified, making it unclear which is the most precise, but it is clear that they follow the same trends throughout the week. Two times, just after 120 hours and 140 hours, our vision system detects more people than the Bluetooth system. This is caused by police cars driving in the pedestrian zone on Friday and Saturday nights. This is not accounted for yet in the vision system. The advantage of a vision system in this application is that it counts the actual number of people, compared to counting a number of, eg. Bluetooth devices, where the relation to the number of people is unknown.

The processing takes 1.2 ms per frame (384×288 pixels), which easily obeys real-time requirements. The test has been performed with a multi-threaded implementation of the system on a laptop with an 2.00GHz Intel Core i7-2630QM CPU and 8GB RAM.

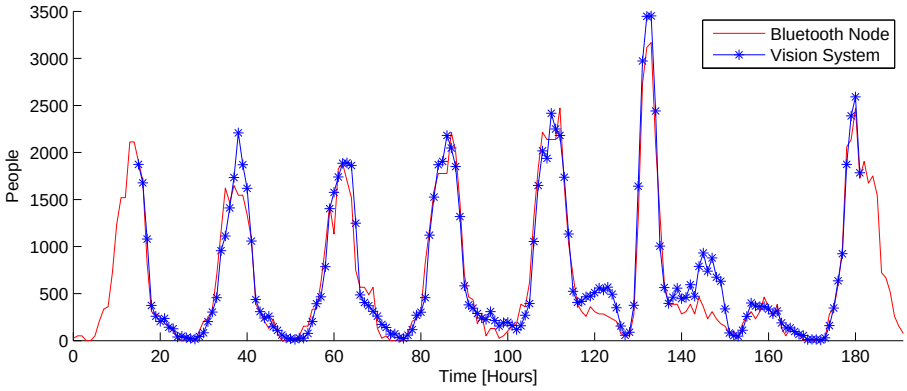


Fig. 11.11: Our system (blue graph) compared to a commercial system based on Bluetooth devices (red graph). The Bluetooth results have been multiplied with a factor 25.76.

11.5 Interactive Urban Lighting

In the coming years, it is expected that the type of urban illumination will change to light emitting diodes (LED) to a greater extent. In addition to being less energy demanding, this type of light source enables digital control of both colour and intensity. That opens up for new possibilities for designing interactive and adaptive lighting. Interactive environments can engage people to feel more connected to the urban spaces and encourage them to stay or to play in the environment. Intelligent lighting can also make the environment appear more comfortable and secure. In this full scale experiment, we put up three thermal cameras and 16 lamps in an urban space. A picture of the space is shown in figure 11.12.

Fig. 11.12: Overview of part of the observed urban space.

Four different light scenarios were tested:

1. Ambient Illumination: Static white lighting.
2. Glowing Light: Each lamp individually fades up and down between 0-20 % intensity.
3. White Aura: Illuminated circle with a diameter of min. 10 meter around each person.
4. Red Treasure Hunt: When a person approaches a "trigger light", a wave of light is sent out through the square.

The two first scenarios do not depend on the movement of people, while the last two are interactive scenarios, where thermal cameras are applied to estimate people's movements [21].

11.5.1 Hardware setup

Figure 11.13 illustrates the hardware of the system. Three thermal cameras are each connected to a computer, which processes the video and converts tracks of each person at the square into a common world coordinate system. This tracking data is sent to a computer that controls the lamps based on the live data and a chosen light scenario. A total of 16 RGB LED lamps are controlled by the system.

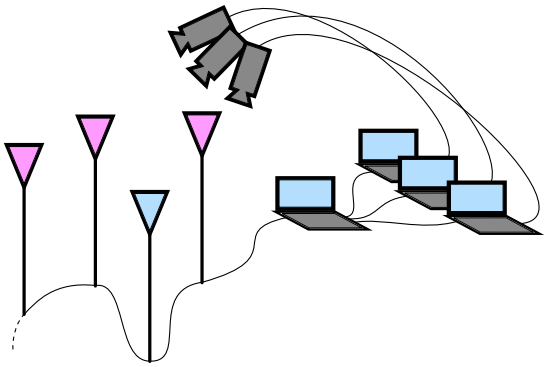


Fig. 11.13: System overview

Inside each lamp is installed a DMX module, which controls each colour of the LED lamp in 255 brightness steps, making it possible to control both colour and brightness of each lamp individually. The area covered by each camera is illustrated in figure 11.14.

The real-time aspects of this project are very important, because the lighting must react to people's concurrent movements and have an update rate that allows for smooth control. Communication between both computers and lamps are handled using the Open Sound Control protocol [22].

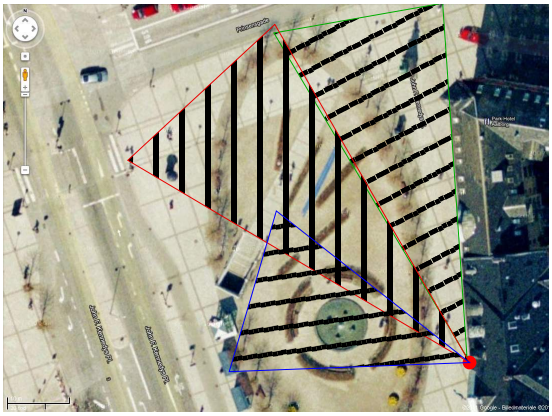


Fig. 11.14: Overview of the square with camera views illustrated.

11.5.2 Methods

Input to the interactive lighting system is the real-time human movement. Since the observed area is an urban space with restricted car access, it is assumed that all observed activity is human activity of interest. People are detected by performing background subtraction, and then the objects are filtered by size. Since the background temperature naturally changes over time, the background must be updated. It is chosen to perform a running average background subtraction with a selective update, meaning that only pixels segmented as background will contribute to the new background. The detected image coordinates of people are transformed into world coordinates using a homography matrix calculated during an initialisation. Further description of the methods can be found in [21]. People walking closer than three metres to each other are treated as groups in this work. Each person or group object is tracked in world coordinates with a Kalman filter [23, 24]. Treating groups of people as one object to track also helps to overcome the problems of occlusions between people, as it is not necessary to distinguish single persons.

The videos from each camera are processed individually, but after transforming the positions, all the tracking and lamp positions refer to the same world coordinate system. The map with overlayed tracking results for a 5-minute period is shown in figure 11.15.

For each frame, the position and velocity vector for each person is registered from the tracking and sent to the light control system. In addition, merging and splitting of groups are registered as special events that could trigger light events.

The reactions from people are analysed both through interviews and from the recorded occupancy and movements.

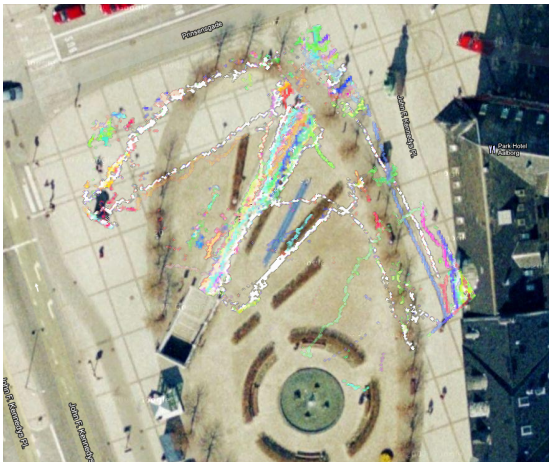


Fig. 11.15: Overview of the square with tracking results overlayed. Each person is assigned one colour.

11.5.3 Results

The system has been tested for one week during the winter, when very cold weather could possibly affect the behaviour of people. Observations and interviews showed that people used the space mostly for transit, and often they did not notice the changed illumination. But people observing the square from outside noticed and appreciated the interactive behaviour [21]. However, later experiments have shown that the lighting causes a range of abnormal behaviour, such as cat walking, dancing and light chasing.

The vision system is evaluated qualitatively by looking at the positions of visualised tracks on a digital map compared to the real position of the person, as well as the actuated lighting scenario compared to the expected scenario. With a precision of approximately one meter for the mapped position, the system works as expected, with correct and fast feedback to the lamps.

The full system runs real-time with a frame rate of 15 fps on 384×288 pixel images with a CPU implementation on a Intel Core i5-2430M 2.4GHz CPU. The real-time performance is crucial here, as the application does not allow delays in the feedback to the lamps. Testing the system for one full week without interruptions prove that the real-time performance is stable during changing conditions. For higher frame rate or larger images the performance could easily be improved using a faster CPU or possibly a multi-threaded implementation.

Figure 11.16 shows two frames from the "White Aura" scenario, where an illuminated circle follows a person.

This study also showed that it is possible to save up to 90 % of the energy for lighting, without people changing their behaviour [21].

The tracking information may both be used for instant control of the lighting as described here, but it could also be used for later evaluation and design of

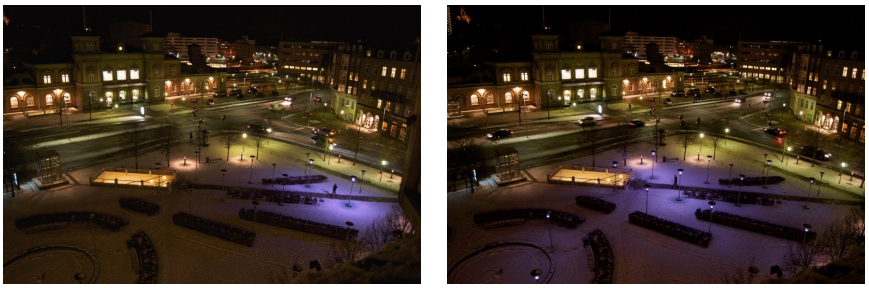


Fig. 11.16: Two frames from the "White Aura" scenario.

urban spaces.

11.6 Automatic Near-collision Detection

In urban environments, the high density of people leads to heavy traffic. Many people tend to use their private cars for comfortable transportation, but in order to reduce emission and congestion, a transition from the car to more eco-friendly means of transportation, such as the bicycle, is needed. To do this, an enhancement of the bicycle conditions must be done in order to increase the share of cyclists. Cities, such as Copenhagen [25], have put special focus on the cyclists, eg., by designing special bike lanes on all the roads. However, studies have shown that bike lanes do not improve the safety of cyclists; even though there is a reduction in the number of crashes on road sections, where bike lanes are established, the number of crashes increases in the intersections; in particular at signalised intersections [26, 27].

In this work, thermal cameras are used to compare different types of intersections in terms of safety. The objective is to evaluate if some geometric designs of bike lanes through intersections are better than others. Since accidents are rare, the evaluation is based on the Swedish Traffic Conflict Technique [28]. The idea is that accidents and near-collisions are related. Thus, the near-collisions will be used as a measure of safety, since they occur more frequently than accidents. The detection of near-collisions can be done manually, but is of course very time consuming and ineffective due to the low number of near-collisions. In the smart city, this can be done automatically using a camera system.

11.6.1 Methods

In this work, we will look for close interactions between cars and bikes, which are prerequisites for near-collisions to happen. Interactions are defined as the simultaneous motion of a car and a bicycle in a given zone of interest. Shown

in figure 11.17 is the car zone in blue and the bike zone in red. When both cars and bikes are present at the same time in their respective zones, there is a risk that the cyclists going straight in the red zone could be hit by a car from lane 1 or lane 2 turning towards lane 3. In both situations, the cars will pass the blue zone with a direction towards lane 3. An example image from one intersection is shown in figure 11.18. Since both position of the camera and the layout of intersections and bike lanes changes between each scene captured, the car and bike zones must be defined manually for each intersection.

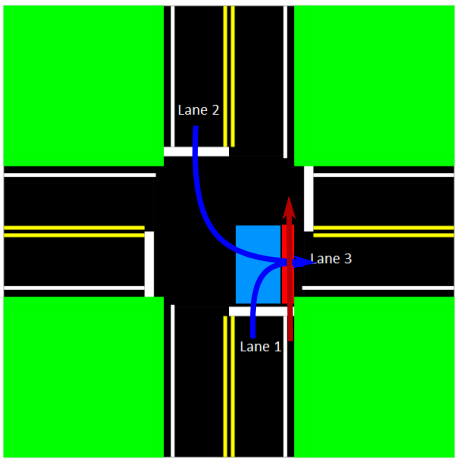


Fig. 11.17: Illustration of an intersection with marked zones of interest. Blue is the car zone, and red is the bike zone. Blue arrows indicate cars' paths of interest, and the red arrow indicates bikes' path of interest.

In order to detect interactions, optical flow [29] is applied to calculate direction and magnitude of the motion. For cyclists, the direction must correspond to going straight, and for cars, they must head towards the cyclist zone. An angle histogram based on the optical flow vectors will be used to decide whether the direction is correct. As shown in figure 11.18 the camera angle is chosen in order to have a straight view on both bike and car lanes, but occlusion between cyclists can occur as they happen to drive close to each other. However, the optical flow method does not rely on detection of individuals, rather the overall motion in the chosen zone. The magnitude of motion must exceed a specified threshold in order to eliminate noise. This threshold depends on the camera location and angle, but due to a relatively small detection area, the effects from image perspective within each image are neglected.

The motion must be consistent for a short interval of time in order to be considered a real detection. Therefore, we implement a buffer that flags the frames in which either a cyclist or car's motion is observed. In order to register an interaction, both buffers must be flagged for a number of consecutive frames.

Fig. 11.18: Example of an interaction between a car and a bicycle.

11.6.2 Results

One hour of video from each of the four intersections is used for testing the system. The videos are captured from 7am to 8am, corresponding to the morning rush hour. All interaction situations are manually labelled and compared to the output of the automatic system. The results show that we get approx. 33 % false positives. This is of course a high rate, but as it is more critical to miss any detections, we allow more false positives. A manual verification process of the detections can be conducted afterwards. A few false negatives are observed. These situations occur when either a car or cyclist has a very low speed, resulting in a motion vector below the threshold.

The speed of the algorithm obeys real-time requirements with a total processing time of 41 ms per frame (640×480 pixels) on a 3.4GHz Intel Core i7-3770 CPU with 8 GB RAM. In idle situations, with no vehicles present, the processing time is 22 ms per frame.

11.7 Analysing the Use of Sports Arenas

The interest in analysing and optimising the use of public facilities in cities has a large variety of applications in both indoor and outdoor spaces. While the previously described projects focus on outdoor spaces, this work copes with indoor scenes. Here we are focusing on sports arenas, but other possible applications could be libraries, museums, shopping malls, etc. We aim to estimate the occupancy of sports arenas in terms of the number of people and their positions in real time. Potential use of this information is both online booking systems, and post-processing of data for analysing the general use of the facilities. For the purpose of analysing the use of the facilities, we also try to estimate the type of sport observed based on people's positions.

11.7.1 Methods

In indoor spaces, the temperature is often kept constant and cooler than the human temperature. Foreground segmentation can therefore be accomplished by automatic thresholding the image. In some cases unwanted hot objects, such as hot water pipes and heaters, can appear in the scene. In these situations, background subtraction can be utilised. After obtaining a binary image, the foreground should be converted to a number of people. Each white blob is simply counted as a person, but in order to handle partial occlusions, the blobs can be split both vertically and horizontally before counting them [30]. As described in section 11.2, no very wide angle lenses exist for thermal cameras yet. To capture a wide area, more cameras can be put together to form a wide angle image. In figure 11.19, three images have been stitched together to cover a 20×40m arena. The lower image shows the segmentation of people by automatic thresholding and removing white pixels outside the region of interest, here the court area.

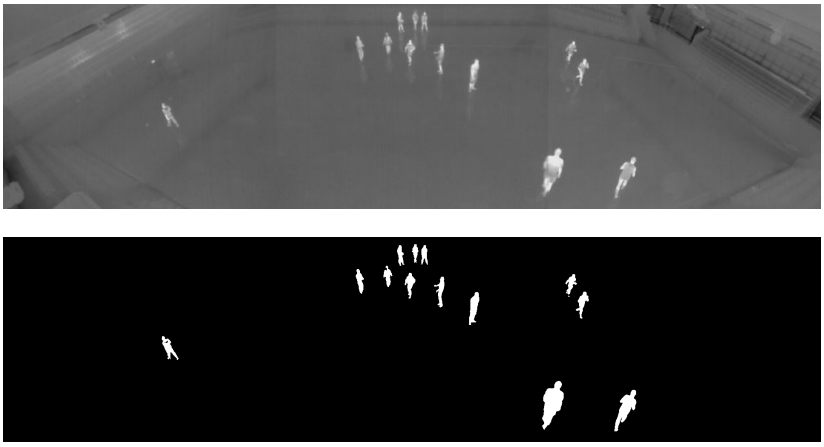
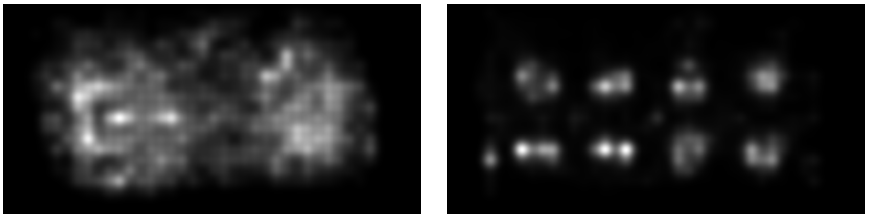


Fig. 11.19: Upper: Thermal input image from a sports arena. Lower: Result after thresholding the image and removing white pixels outside the court area (ROI).

For crowded scenes, occlusions will result in missed detections, and reflections and non-human objects can result in false detections as well. Including temporal information may stabilise these measurements. Detecting when people leave or enter the scene gives information about the transitions in number; during periods with no activity in the border area, the number of people inside the monitored area must be the same. Using a dynamic programming approach, the results are optimised over long periods for more stable measurements [31].

From the detection of people, their activities can also be analysed. For sports arenas, we estimate the sports type from the occupancy patterns over 10-minute periods. Each registered position of people on the court contributes with a Gaussian distribution to a heatmap. An example of two heatmaps for

basketball and badminton are shown in figure 11.20.



(a) Basketball (b) Badminton

Fig. 11.20: Heatmaps for 10-minute periods.

A classifier is trained to detect five different sports types from the heatmaps only. From these images, the dimensionality is reduced with PCA and Fisher's Linear Discriminant (FLD), where FLD uses labelled training data and seeks the dimensions which discriminate the classes. After transforming each sample data to the new space, the nearest class is found using the Euclidean distance [32].

11.7.2 Results

The detecting method has an error rate between 7.35-11.76%, depending on the activity level in the videos. Adding the border activity detection and dynamic programming optimisation reduces the error rate to 4.44 %. This full approach is tested on 30 minutes of annotated video, while the underlying detection method has been tested on a large annotated dataset from five different arenas. The detection algorithm easily runs real-time, with a processing time of 12.5 ms per frame for large images of 1920×480 pixels. The approach described in [31], which includes tracking of people near the border takes 60ms per frame, without any optimisation of the software. This means that even this method will be able to perform in real-time. However, the dynamic programming optimisation is a post-processing method and must run after the full period of video has been processed. Both methods are tested on an Intel Core i7-3770K 3.5 GHz CPU with 8GB RAM, and processing 1920×480 pixel images.

The sports type classification has been tested on 30 heatmaps representing five different sports types, plus a category of miscellaneous activities. The recognition rate is 90.76%.

11.8 Mapping and Modelling Human Movement and Behaviour in Public Spaces

In matters of urban planning and management, it is essential to know how streets and public spaces are being used and how people move around. To

quantify and eventually model human movements and patterns of use of a public space over time, which we refer to as Human Spatial Dynamics (HSD), it is thus necessary to track each individual crossing of the space under scrutiny. Pivotal to this research is the use of Geographical Information Systems (GIS). The idea is to use computer vision technology to extract accurate geo-referenced tracks of people and use GIS based methods to store and analyse the HSD data created. The advantage of utilising GIS is that the HSD data captured can be easily related to other geospatial data layers and be directly available in the GIS workflow of professional planners and managers.

Along with the advances in software technology and computing power in the last decade, there has been a growing interest in modelling pedestrian and bicyclist behaviour based on a bottom-up approach of programming them as individual entities or agents, which can interact in simulations and yield emergent movement patterns and behaviours resembling those observed in the real world. Collectively, these micro models are referred to as Agent-Based Models (ABM) and there are several approaches to programme the underlying principles. Concepts such as Social Forces [33, 34], Cellular Automata [35], Behavioural Heuristics [36, 37], Discrete Choice [38–40] and Behavioral Geography and spatial cognition [41] have been suggested. However, most models focus specifically on crowd or evacuation dynamics and not so much on modelling entire trips of pedestrians in regular traffic in public spaces. Despite advances in modelling techniques towards sophisticated ABMs, they still have challenges in reproducing real world behaviours reliably in all situations [42–44]. A main reason for that collectively mentioned in the literature is that there exists few empirical studies and verified standard HSD datasets from recordings of real life pedestrian and bicycle traffic to calibrate the models against. The thermal video tracking technology holds the potential for being a way to collect long time HSD datasets in various places, and thus contribute to improve the ABM models. This quantitative approach to tracking urban public life may also be able to supplement the traditional and intuitive manual approaches to HSD data collection used in the studies of urban public spaces and qualities. Inspired by the works of [45] and [46], a possible outcome of the project is also to contribute with new digital methods to this field.

11.8.1 Methods

A pilot study was made to prove the concept. A pedestrian zone in Copenhagen with occasional bicycle traffic and goods delivery by vehicles was used as a test scene. The scene was situated where one of the city's major shopping streets meets a perpendicular street and an open square on the way to a major subway station. The site therefore had a continuous flow of pedestrians from several directions that had to negotiate with others to make their way through the scene. At the same time, there were also people in the scene waiting, meeting and talking for longer periods of time. People dragging their bikes or pushing prams or wheelchairs were also observed, as well as groups of school children on

excursions. Occasionally, cyclists were observed riding their bikes despite the legislation. Figure 11.4 shows the scene as both an RGB and thermal image, though with slightly different views. The camera was placed as high as possible on the roof top terrace of a 5 story building next to the scene to minimise the number of people occluding others in the camera FOV, while at the same time capturing the traffic of as large an area as possible. Control points in the scene used to calibrate the homography matrix transferring image pixel coordinates to real world coordinates were measured with high precision GPS equipment.

Computer vision software was applied to analyse 5 minutes of thermal video for which a simultaneous RGB video was also recorded for reference. Background subtraction was used in order to detect people. Since the observed scene was very busy, it was not possible to find an empty frame which could be used as background. Instead, a background model was obtained by calculating the median value for each pixel over a 30 second initialising period. The background was updated during run-time, using a selective update method equal to the one described in section 11.5.2. The foreground objects were filtered by size, in order to remove noise. Even though the camera was mounted at a high position, occlusions still caused problems due to the high density of people. In order to solve partial occlusions, we are able to split blobs both vertically and horizontally [30]. After converting the position of the remaining objects to world coordinates, they were tracked using Kalman filtering [23]. For each frame, the tracking software yields a list of ID numbers and positions of the detected persons in real world coordinates. To read the data in a GIS, the raw files were passed by a Python script to render a list of locations for the tracking of each of the IDs. To reduce the amount of data processed while still maintaining sufficient location accuracy the points were down-sampled to fewer instances per second than the original frame rate (30 fps). During parsing of the raw files, a series of attributes were added to the individual points, including speed (in relation to the previous point, a given interval back in time, and accumulated for the track up to the given point) and incremental distance and time. Further metadata was generated for each individual track, including distance, duration, Euclidean distance, average speed, number of points etc.

11.8.2 Results

The attributes added allowed for various ways of sorting and visualizing the tracks in GIS. First noise and false detection needed to be identified and removed. A threshold value of 3 seconds was chosen as the acceptable quality criterion for the minimum duration of a track. From the segmentation of the thermal video described in figure 11.6 it was known that there were areas with a high degree of noise, particularly to the right in the FOV near the camera, caused by wind movement of the edges of some sun shades in the scene. Tracks detected along or intersecting the edges of this area were thus classified as unreliable and removed from the rest of the tracks. In figure 11.21 the green lines depict the tracks that met the quality criteria (460 tracks) whereas the

red crosses indicate detections that were too short in time or distance (1475 IDs). The green tracks clearly show the movement patterns of the area as well as the density of the tracks indicating which routes were the most used. The shaded areas depict the obstacles that the traffic had to evade. An interesting observation concerning the obstacles is the diverging traffic flow in the upper part of the figure, where it is clearly seen that tracks split into two routes, indicating that people are passing either to the left or right around the shaded area in the urban square seen outside the FOV.

Concerning the false detections, the ones caused by the sun shades are clearly seen in the FOV to the right of the camera position. A dense cluster of short detections is also seen near the camera. This was caused by two persons standing close together at that same spot talking for the entire period analysed, including the initialisation period. This caused the two persons to be part of the background model, but small movements and gestures of the people generate short detections. A dense cluster of short detections is also evident at the entrance to the building in the southern end of the FOV. This was caused by several persons passing the doorway simultaneously, thus creating heavy occlusions. Several short detections are also seen spread out across the area where the green tracks dominate. These are probably caused by people occluding each other when walking close together or passing others in the scene. The number of occlusions would probably be able to be lowered substantially with a nadir looking camera to monitor the scene.

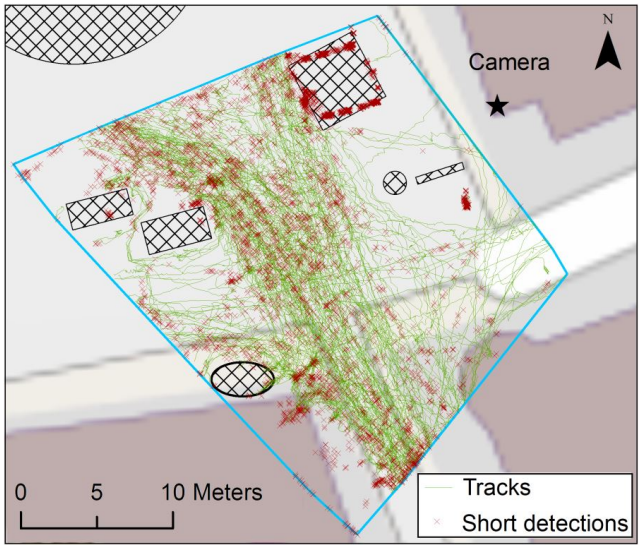


Fig. 11.21: The green lines indicate all the tracks from the 5 minute period analysed, where an ID was followed for more than 3 seconds. The red crosses represent short detections that did not fit the quality criterion. The shaded areas represent obstacles that the traffic had to evade.

The GIS setup made it possible to make analysis of each individual track and to compare tracks spatio-temporally in order to assess behavioural patterns and the HSD seen in the scene. To show an example of this, four tracks of people moving in the same direction in the same 30 second period were selected and colourized according to their speed as shown in figure 11.22. Both increasing and decreasing speeds are seen on the tracks.

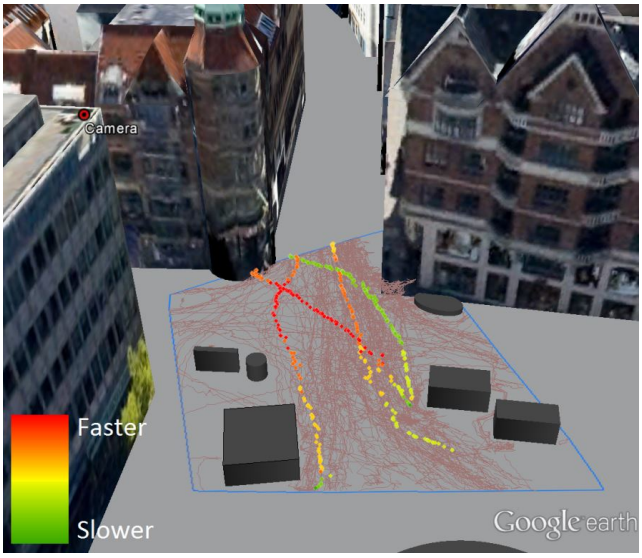


Fig. 11.22: The site and the FOV illustrated in a 3D view with four selected tracks from the same 30 second period coloured according to their speed. All four tracks start in the bottom of the image in the open square and move up towards the street. The tracks indicate both increasing and decreasing speeds. All tracks from figure 11.21 are displayed as background. The position of the camera is shown on the corner of the building to the left. The obstacles are indicated in grey.

To assess the method's overall ability to measure speed, the distribution of the average speeds of all tracks from figure 11.21 were plotted in the histogram shown in figure 11.23. The graph shows a normal distribution of speeds around 5 km/h, which is typical for pedestrian traffic. In the lower extreme around speeds of 1-2 km/h there are more tracks than in the higher end around 8-9 km/h. This is possibly due the fact that more people are stopping or waiting briefly while in the scene, which lowers their average speed, as opposed to fewer individuals that hurry through the scene. The few tracks in the upper extreme around 13-14km/h were identified as bicyclist in the video.

Further research based on this project will aim to develop more advanced GIS methods to study behaviour, such as people's choices of direction and speed, and the interaction with others, in order to enable extraction of behavioural parameters that can be used in ABMs. The analyses shown here were all made as post-processing procedures, but there is nothing hindering

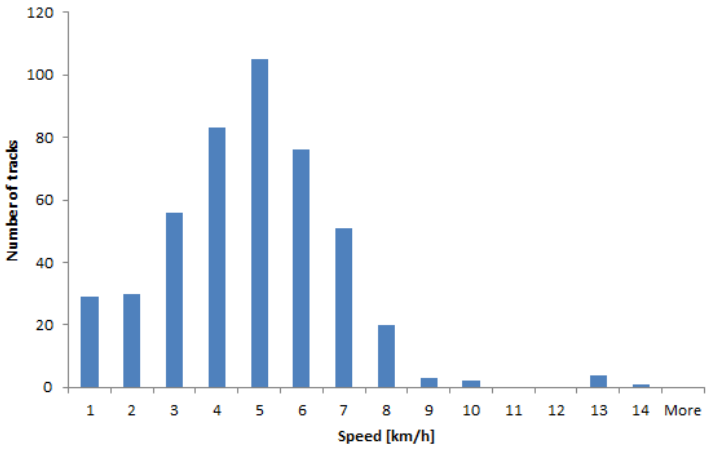


Fig. 11.23: The graph shows the distribution of average speeds of the green tracks displayed in figure 11.21. The distribution is as expected for pedestrians with speeds normally distributed around 5 km/h. The few tracks with an average speed of around 13 km/h were identified as bicyclists in the video.

the GIS analysis from being automated to generate near real-time online maps of the tracked scene. The processing time of the computer vision tracking algorithm was 20 ms per frame for 640×480 pixel images on an Intel Core i7-3770K 3.5 GHz CPU with 8GB RAM. Even without any optimisation or parallelization of the algorithm this easily obey real-time requirements, and could be used as input for any real-time analysis of the human behaviour in the public space. GIS methods could thus also be applied in conjunction with some of the other projects presented to make spatial analyses of the tracks generated.

11.9 Conclusion

The thermal camera is considered an important technology for use in the future Smart Cities. Being a non-intrusive passive sensor, which also preserves privacy, makes it very suitable for the purpose. Furthermore, the thermal camera is independent of light, and thereby operates equally well during day and night, compared to other sensors like the RGB camera; which is strongly dependent on sufficient and stable lighting. For the purposes of e.g. people counting and simple tracking, thermal imaging is highly suited. Some segmentation methods, such as thresholding and image differencing are extremely fast, and still accurate. For more complex tasks, such as tracking of individual people through the city, a different technology able to detect unique features or ID's must be applied.

This paper presented five different Smart City applications in which we applied thermal imaging. They cover both indoor and outdoor environments,

monitoring the movements of people, cars and bikes. All systems have proven to be real-time compatible and are tested over very long time in real-world settings.

We have shown here, that by employing thermal cameras it is possible to measure the human use of a city, without violating the privacy of citizens. For the expected future scenario of large scale implementation of intelligent technology in smart cities, we find it crucial to consider sensors and methods that protect the privacy of people. Furthermore, being able to operate day and night without any manual involvement opens up a great number of new applications. Thus, the applications presented in this paper could easily be extended to other smart city applications based on detection and tracking of humans or vehicles.

Acknowledgement

The authors direct a special thanks to Nordea-fonden, The Danish Foundation for Culture and Sports Facilities, The Danish Road Directorate and Danish Lighting Innovation Network for economical support. We would also like to thank the following partners for an innovative collaboration: Teamtronic A/S, Riegens A/S, Alfred Priess A/S and Lund University, and we would like to thank the administration of the building Knud Højgaards Hus and the company Fokustranslatørerne for allowing access to their roof top terrace for the recording of data.

References

[1] M. Batty, K. Axhausen, F. Gianotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, and Y. Portugali, "Smart cities of the future," *Eur. Phys. J. Special Topics*, vol. 214, pp. 481–518, 2012.

[2] Copenhagen Cleantech Cluster. (2012) Danish smart cities: Sustainable living in an urban world. [Online]. Available: http://www.cphcleantech.com/home/publications/\reports/smart-city-report-2012

[3] O. B. Jensen, *Staging Mobilities*, 1st ed. Routledge, 2013.

[4] M. Shepard (ed.), *Sentient City: Ubiquitous computing, architecture, and the future of urban space.* MIT press, 2011.

[5] R. Kitchin, "The programmable city," *Environment and Planning B: Planning and Design*, vol. 38, no. 6, pp. 945–951, 2011.

[6] O. B. Jensen, "The Will to Connection - a research agenda for the 'programmable city' and an ICT 'toolbox' for urban planning," in *Local and Mobile: Linking Mobilities, Mobile Communication, and Locative Media.* Routledge, 2013.

[7] Panasonic Electric Works Corp. (2013) Infrared Array Sensor: Grid-EYE. [Online]. Available: http://pewa.panasonic.com/components/built-in-sensors/infrared-array-sensors/grid-eye/

[8] FLIR Systems Inc. (2013) FLIR X8400 sc specifications. [Online]. Available: http://flir.com/cs/emea/en/view/?id=52430

[9] ——. (2013) FLIR T620 & T640 datasheet. [Online]. Available: http://www.flir.com/uploadedFiles/\Thermography_USA/Products/Product_Literature/\flir-t620-datasheet.pdf

[10] AXIS Communications. (2013) Axis network cameras. [Online]. Available: http://www.axis.com/products/video/camera/\index.htm

[11] ——. (2013) Axis IP surveillance. [Online]. Available: http://www.axis.com/files/brochure/bc_ipsurv_49693_en_1211_lo.pdf

[12] K. Al-Mutib, M. Emaduddin, M. AlSulaiman, H. Ramdane, and E. Mattar, "Motion periodicity-based pedestrian detection and particle filter-based pedestrian tracking using stereo vision camera," *Int. J. Comput. Appl. Technol.*, vol. 50, no. 1/2, pp. 113–121, Jul. 2014.

[13] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008, pp. 1–8.

[14] T. B. Moeslund, A. Hilton, V. Krüger, and L. Sigal, Eds., *Visual Analysis of Humans - Looking at People*. Springer, 2011.

[15] J. Kapur, P. Sahoo, and A. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Computer Vision, Graphics, and Image Processing*, vol. 29, no. 3, pp. 273 – 285, 1985.

[16] S. Ouadfel and S. Meshoul, "Bio-inspired algorithms for multilevel image thresholding," *Int. J. Comput. Appl. Technol.*, vol. 49, no. 3/4, pp. 207–226, Jun. 2014.

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 886–893 vol. 1.

[18] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[19] B. Shoushtarian and H. E. Bez, "A practical adaptive approach for dynamic background subtraction using an invariant colour model and object tracking," *Pattern Recognition Letters*, vol. 26, no. 1, pp. 5 – 26, 2005.

[20] BLIP Systems A/S. (2013) BLIP Systems A/S. [Online]. Available: http://www.blipsystems.com/

[21] E. S. Poulsen, H. J. Andersen, O. B. Jensen, R. Gade, T. Thyrrestrup, and T. B. Moeslund, "Controlling urban lighting by human motion patterns - results from a full scale experiment," in *ACM International Conference on Multimedia (MM)*, 2012.

[22] The Center For New Music and Audio Technology (CNMAT), UC Berkeley. (2013) Open sound control. [Online]. Available: http://opensoundcontrol.org

[23] G. Welch and G. Bishop, "An introduction to the kalman filter," Chapel Hill, NC, USA, Tech. Rep., 1995.

[24] A. Ali and K. Terada, "A fast approach for person detection and tracking," *Int. J. Comput. Appl. Technol.*, vol. 44, no. 3, pp. 210–216, Sep. 2012.

[25] T. A. S. Nielsen, H. Skov-Petersen, and T. Agervig Carstensen, "Urban planning practices for bikeable cities - the case of Copenhagen," *Urban Research & Practice*, vol. 6, no. 1, pp. 110–115, 2013.

[26] N. Agerholm, S. Caspersen, J. Madsen, and H. Lahrmann, "Cykelstiers trafiksikkerhed; en før-efterundersøgelse af 46 nye cykelstiers sikkerhedsmæssige effekt [traffic safety of bike lanes; a before-and-after investigation of the safety effect of 46 new bike lanes]," *Dansk Vejtidsskrift*, vol. 83, no. 12, pp. 52–57, 2006.

[27] S. Jensen, *Effekter af Cykelstier og Cykelbaner: Før-og-efter evaluering af trafiksikkerhed og trafikmængder ved anlæg af ensrettede cykelstier og cykelbaner i Københavns Kommune [The effects of bike paths and bike lanes; Before-and-after evaluation of traffic safety and traffic volume when establishing one-way bike paths and bike lanes i Copenhagen Municipality]*. Trafitec, 2006.

[28] C. Hyden, *The development of a method for traffic safety evaluation: The Swedish Traffic Conflicts Technique*. Department of Traffic Planning and Engineering, Lund Institute of Technology, Sweden, 1987.

[29] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[30] R. Gade, A. Jørgensen, and T. B. Moeslund, "Occupancy analysis of sports arenas using thermal imaging," in *Proceedings of the International Conference on Computer Vision and Applications*, 2012.

[31] ——, "Long-term occupancy analysis using graph-based optimisation in thermal imagery," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[32] R. Gade and T. B. Moeslund, "Sports type classification using signature heatmaps," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013.

[33] D. Helbing, P. Molnár, I. J. Farkas, and K. Bolay, "Self-organizing pedestrian movement," *Environment and Planning B: Planning and Design*, vol. 28, no. 3, pp. 361–383, 2001.

[34] M. Moussaïd, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz, "The walking behaviour of pedestrian social groups and its impact on crowd dynamics," *PloS one*, vol. 5, no. 4, 2010.

[35] V. J. Blue and J. L. Adler, "Cellular automata microsimulation for modeling bi-directional pedestrian walkways," *Transportation Research Part B: Methodological*, vol. 35, no. 3, pp. 293 – 312, 2001.

[36] M. Moussaïd, D. Helbing, S. Garnier, A. Johansson, M. Combe, and G. Theraulaz, "Experimental study of the behavioural mechanisms underlying self-organization in human crowds," *Proceedings of the Royal Society B: Biological Sciences*, vol. 276, no. 1668, pp. 2755–2762, 2009.

[37] M. Moussaïd, D. Helbing, and G. Theraulaz, "How simple rules determine pedestrian behavior and crowd disasters," *Proceedings of the National Academy of Sciences*, vol. 108, no. 17, pp. 6884–6888, 2011.

[38] G. Antonini, M. Bierlaire, and M. Weber, "Discrete choice models of pedestrian walking behavior," *Transportation Research Part B: Methodological*, vol. 40, no. 8, pp. 667 – 687, 2006.

[39] M. Bierlaire and T. Robin, "Pedestrians choices," in *Pedestrian Behavior. Models, Data Collection and Applications*, H. Timmermans, Ed. Emerald Group Publishing Limited, 2009, pp. 1–26.

[40] T. Robin, G. Antonini, M. Bierlaire, and J. Cruz, "Specification, estimation and validation of a pedestrian walking behavior model," *Transportation Research Part B: Methodological*, vol. 43, no. 1, pp. 36 – 56, 2009.

[41] P. M. Torrens, "Moving agent pedestrians through space and time," *Annals of the Association of American Geographers*, vol. 102, no. 1, pp. 35–66, 2012.

[42] C. J. E. Castle and A. T. Crooks, "Principles and concepts of agent-based modelling for developing geospatial simulations," *CASA working papers*, 2006.

[43] A. Crooks, C. Castle, and M. Batty, "Key challenges in agent-based modelling for geo-spatial simulation," *Computers, Environment and Urban Systems*, vol. 32, no. 6, pp. 417 – 430, 2008.

[44] E. Papadimitriou, G. Yannis, and J. Golias, "A critical assessment of pedestrian behaviour models," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 12, no. 3, pp. 242 – 255, 2009.

[45] W. H. Whyte, *The Social Life of Small Urban Spaces*. Project for Public Spaces Inc, 1980.

[46] J. Gehl, *Cities for People*. Island Press, 2010.

Chapter 12

Controlling Urban Lighting by Human Motion Patterns - Results from a Full Scale Experiment

Esben Skouboe Poulsen, Hans Jørgen Andersen, Ole B. Jensen,
Rikke Gade, Tobias Thyrrestrup and Thomas B. Moeslund

The paper has been published in
Proceedings of the ACM International Conference on Multimedia,
pp. 339–347, October 2012.

© 2014 ACM
The layout has been revised.

Abstract

This paper presents a full-scale experiment investigating the use of human motion intensities as input for interactive illumination of a town square in the city of Aalborg in Denmark. As illuminators sixteen 3.5 meter high RGB LED lamps were used. The activity on the square was monitored by three thermal cameras and analysed by computer vision software from which motion intensity maps and peoples trajectories were estimated and used as input to control the interactive illumination. The paper introduces a 2-layered interactive light strategy addressing ambient and effect illumination criteria totally four light scenarios were designed and tested. The result shows that in general people immersed in the street lighting did not notice that the light changed according to their presence or actions, but people watching from the edge of the square noticed the interaction between the illumination and the immersed persons. The experiment also demonstrated that interactive can give significant power savings. In the current experiment there was a difference of 92% between the most and less energy consuming light scenario.

12.1 Introduction

For the first time in human history, more than half of the human population inhabits urban environments, and this now presents itself as second nature to humans. The urban context is of a very complex nature and is composed by a multitude of different networks, infrastructures and volumes. The city has become the dominant scenery for everyday life. As such it presents still greater design challenges for an improved urban spatial performance, which can adapt to changes and present inspiring, efficient and stimulating public spaces.

One must acknowledge that urban spaces are sites of movement and interaction that contain unused potential [1, 2]. If we can monitor and potentially understand how the urban space is used in terms of movement and occupancy patterns, we can generate site specific maps that can be used to control elements in the environment such as the illumination.

People will in this way interact direct or indirect with elements in the environments, thus establish an exchange also described as feedback in the world of computation. The study of feedback in computer systems, describing the relationship between sensor and acting environments, can be tracked back to Norbert Wieners [3] notion of cybernetics; a marriage of control theory, information science, and biology that seeks to explain the common principles of control and communication in both animals and machines [3]. Since then much work has been done in the field of computer human interaction. Research fields such as robotics [4], responsive environments [5], situated technologies [6] all contributes to a particular focus within the field of sensing and responding to change.

Inspired by the cybernetic notions from Gorden Pask [7] Usman Haque

(a) Overview of the Kennedy square without the lamps, seen from the position of the thermal cameras.

(b) Kennedy square with the lamps at daytime. The main train station in the background.

Fig. 12.1: Kennedy square in the city of Aalborg in Denmark.

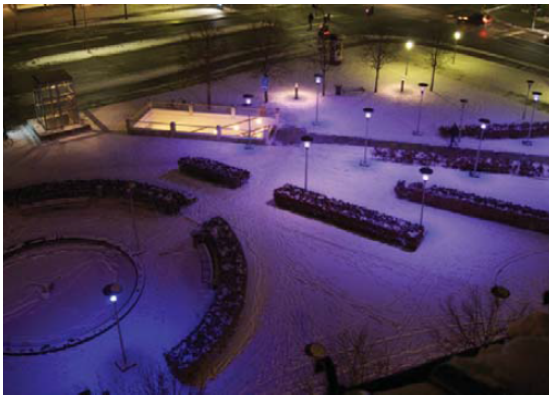


Fig. 12.2: Overview of the Kennedy square with the 16 lamps, seen from the position of the thermal cameras.

presents experiments that utilize sensor technologies and lighting as part of a larger collective constructed environment where people and objects collaboratively create social domains as in the case of Sky Ear and Open Burble [8]. In the two cases environmental feedback between weather systems (electromagnetic waves and local winds) and social actors are essential in the temporary composition of the color and the intensity of the light. Hence the balloons mediate a direct relation between subjects and weather phenomena, it establish a platform for conversations, between the humans, calling a balloon, causing waves of light to appear between the sky ear and the observers. On the level of the subject, internal conversations presents questions related to the level of engagement; to what level does I feel like participating? As performer, actor or just a passive observer enjoying the show? More external stimuli are constantly affecting our choices and engagement in the event, (weather, social

norms, economy etc.) these internal and external conversations is a keystone in the understanding of systems and knowledge produced within them [9]. Another example is MY Studios Low Rez reactive sound poles [10], which is a spatial sequencer, allowing the by-passer to compose different sound-scapes. This installation has proven to be a mediator for vivid interactions between performers and observers of the event [11].

Within the contemporary art field Philipe Beesley [12] pushes the exploration of responsive environment inspired by reaction patterns in nature, exemplified in the art installations Endothelium and Hydrozoic. In these Beesley explore how he, inspired by nature, can design artificial life forms that produce life like behaviors and appearances, this artistic approach motivate a series of internal conversations; What is this? Is it a life? I am just a significant small part of the interaction pattern, but still I affect the ecology like behavior etc. Thus the installations are not build to be interactive toys but rather responsive environments, where relations and interdependencies are to be explored. In the project Dune [13] the artist Daan Roosegaarde brings a 60 meter long permanent reactive light and sound installation into the landscape along the Mass river in Rotterdam. The installation consists of thousands illuminating light straws that react in different behaviors. Similar to Beeslys the behavior of the lighting is inspired by natural mechanisms, hence it can be scared, excited, curios. Within the last decade these and many other art installations represent cases in the emergence of a new artistic expression in media, interactive and installation art that has proven a creative aesthetic, social and architectural potential for a design discipline concerned with feedback between humans and environments.

Within the industry of lighting, firms like Echelon [14] and Philips [15] has entered the development of technologies for large-scale control systems of outdoor lighting. They contribute to the development of tools and techniques for new types of light designs, which open the window into the design of behaviors of the light. Within the larger urban perspective notions of the smart grid [16] also support this trend of a more efficient and intelligent energy use. This responsive paradigm is deeply related to the discipline of design.

This study will focus on the development, implementation and investigation of interactive urban lightings. It is motivated by the fact that in the coming years urban illumination will change from light sources based on mercury or similar technologies into Light Emitting Diodes (LED) or other less energy demanding light sources. Especially LED light sources is from a control point of view interesting as they will open for new ways to control the illumination, allowing designers and engineers to develope interactive light settings that for example dim or change color in real-time.

To fully release this potential it will be necessary to automatically monitor the activity in the urban environments. This task is very difficult and can be solved at various levels of detail. In the fully monitored situation a persons status of mind could be estimated, as being glad, sad, busy, relax etc. This is, however, at the moment not feasible with the current sensing technologies and

models for human behavior at hand, which will require front views and high resolution images of the face [17, 18]. However, computer vision technologies has developed so it is now possible to track peoples trajectory in urban environments and in this context especially newer thermal imaging cameras has shown promising results [19]. In this study we will utilize recent results within thermal computer vision combined with new LED lights for development of an interactive illumination setting where the ambition is to make the illumination provide safety, being energy effective, aesthetic appealing and potentially give the people another experience of the urban space.

During January these technologies were used in a full-scale experiment on a town square in the city of Aalborg in Denmark. Figure 12.1(a) shows an overview of the square seen from the position of the thermal cameras and figure 12.1(b) shows the interactive lamps at daytime. In figure 12.2 the light setup is shown at nighttime. The activity on the square was monitored using three thermal cameras mounted on a building facing the square. The trajectories of the pedestrians where estimated using computer vision analysis and used as input to control the light behavior in four different responsive light scenarios.

The finding of the study was that in general people immersed in the interactive light setting did not notice that the illumination was changing according to their motion despite the fact that in one of the scenarios the illumination was dimmed significantly in comparison to the normal illumination of the square. On the other hand people watching the square from the outside clearly noticed the interactive light setting and hence got a new experience and impression of the square.

The paper is organised as follows, first the experimental site and setup is presented followed by and introducing to the applied computer vision analysis and the interactive illumination design. The observations of the four different light scenarios are then presented and discussed and finally the findings are concluded.

12.2 Material and Methods

The experiment took place at Kennedy square in the city of Aalborg in Denmark. The square is located between the main train and bus station and the city center and serves primarily as a transit space between these two locations, see figure 1(a) To monitor the square three thermal cameras type Axis Q-1921-E, with a 19mm lens were mounted in the height of 15 meters at one of the buildings facing the square. The cameras covered the area from where an exit of a subway from the main train station is located to where people leave the square on their way to the city center, as illustrated in figure 12.3.

Along this pathway the 16 RGB LED lamps were placed, figure 12.1. The street lamp is composed of a 70 cm tall Riegens Ray light fixture, figure 12.4(a), a 3.5 meter tall light post and a 60x60 cm concrete tile as foundation. An LED module containing 18 1W LEDs, six in each color (RGB) was mounted in the

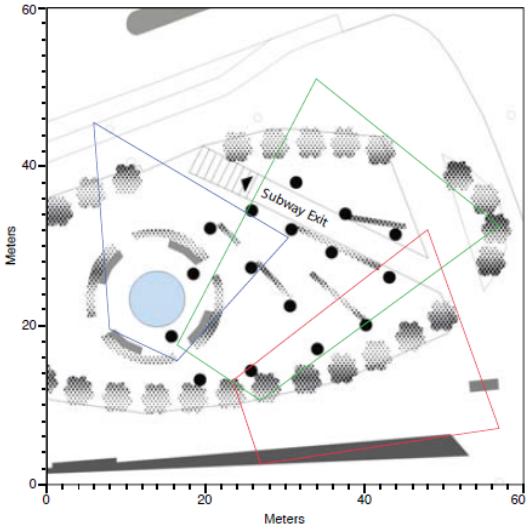


Fig. 12.3: Overview map of Kennedy square with the areas covered by the three cameras.

bottom of the light fixture. The individual LEDs are mounted on the module in a two rings configuration, where the inner ring holds the 6 red LEDs and the outer ring holds 6 green LEDs interleaved with 6 blue LEDs, figure 12.4(b). The LED module is connected to a DMX module installed inside the lamp post. This module enables a 0-255 step brightness control of each led color as well as an unique address of each lamp.

(a) Profile of the lamp housing. (b) LED module.

Fig. 12.4: The LED lamp head.

12.2.1 Qualitative assessment methodology

To evaluate the performance of the computer vision analysis and the control of the illumination a mobile phone application was used to display the activity at the square and the light setting of the lamps, figure 12.5. The application also

gave the possibility to control the lamps. Using the application it was possible in real-time to evaluate if the computer vision analysis worked as expected and that the light scenarios gave the right response according to the position and velocity of people passing through the light setup.

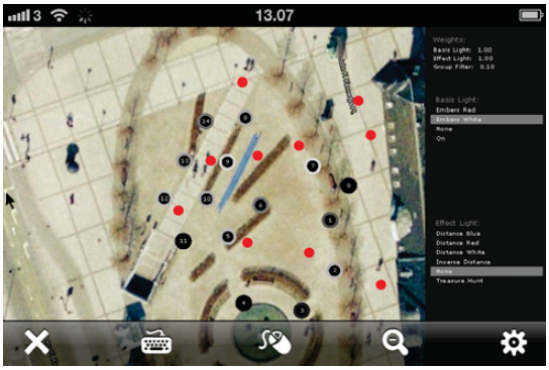


Fig. 12.5: Snapshot of the iPhone application by which it was possible in real-time to see the result of the computer vision tracking and the resulting control of lamps. It was also possible to manually control the lamps by the application.

12.2.2 Observations

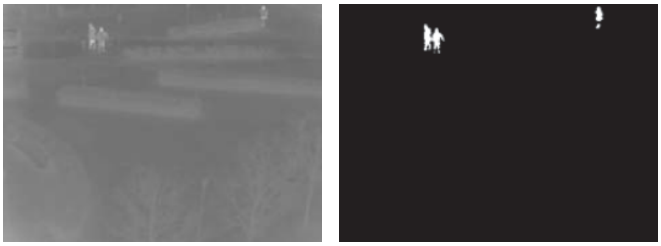
During the experiment observations of body language, gestures and behavior of the occupants on the square were made. The observation methodology was based on ethnographic studies and fieldwork techniques as they are articulated by the sociologists such as Erving Goffman and Edward T. Hall [2, 20, 21] and utilized by architects such as Jan Gehl, William Whyte [22, 23]. By observing the interactions from the edge of the street we were able to describe "space routines" in the transit space [22] and evaluate if people where immersed in or affected by the different light scenarios.

12.2.3 Computer vision analysis

Detecting and tracking people is a large research area in computer vision, most approaches using normal visual cameras. But due to falling prices on thermal cameras new approaches using these sensors have been developed lately. Thermal cameras measure the amount of thermal radiation that lies in the long-wavelength infrared spectrum (8-15 μm). All objects with a temperature higher than the absolute zero emit thermal radiation. The intensity and dominating wavelength depends on the temperature.

Since thermal cameras do not measure visible light, they have a clear advantage over visual cameras in night conditions. Also with visual cameras it is illegal to film public places in Denmark, but since people can not be identified from thermal images, this is not an issue with thermal cameras. This is the

primary reason to utilize thermal cameras in this work. Figure 12.6(a) shows an example of the input image from one of the camera views.



(a) Thermal image. (b) Binary image.

Fig. 12.6: Example of the thermal input image and resulting binary image from one camera view.

As input to the illumination system real-time information about the position and velocity of people at the square should be found. The experiment is conducted in outdoor environment running continuously for a week. The temperature naturally changes, which gives a slowly changing background. Therefore it is chosen to perform a running average background subtraction as the first step in detecting people [24]. The background will be updated with selectivity, meaning that only if the pixel is segmented as background it will contribute to the new background:

$$B_{t+1}(x,y) = \begin{cases} \alpha F_t(x,y) + (1-\alpha) B_t(x,y) & \text{if } F_t(x,y) \text{ background} \\ B_t(x,y) & \text{if } F_t(x,y) \text{ foreground} \end{cases} \tag{12.1}$$

where B_t is the current background, F_t the current input image, α the learning rate and B_{t+1} the new background image. As the experiments take place at an urban space with limited car access, it is assumed that all activity detected is human activity of interest. The difference image produced by background subtraction is binarised with a threshold value of 10. The resulting BLOB's with area of minimum 10 pixels are considered objects of interest. The detected objects must be mapped to real world coordinates at the square in order to correspond to the lamps positions. The mapping is calculated using a homography matrix [25]. This matrix can be calculated using at least four corresponding points in image and world coordinates. Since the camera views are not aligned, a mapping must be calculated for each individual view. Figure 12.7 illustrates the mapping from image to world plane.

For this purpose of illumination control from tracking data, groups of people should be considered as one object. The individual objects detected are grouped using single linkage clustering [26] with a distance threshold of 3 meters. This grouping also ensures that an imprecise segmentation, resulting in one person split in more BLOB's, or more persons found in one BLOB, does not change

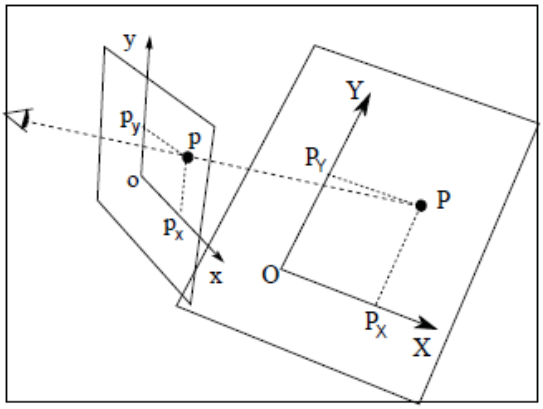
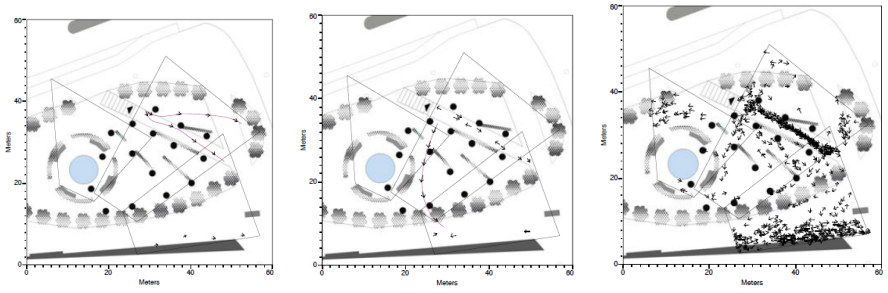


Fig. 12.7: Illustration of the mapping from image to world plane.

the result, as they will be treated as one group anyway. In order to determine stable positions and velocity of the groups, a Kalman filter is applied to track the groups [27]. Splits and merging of groups are handled based on the distance between predictions of group tracks.



(a) One minute trajectories for two persons splitting up and leaving the subway.

(b) Two minute trajectories, with one person showing a distinct path towards the Bank located on the square.

(c) One hour velocity map of the activity that clearly demonstrates the used of the square. Especially, the use of the pathway from the subway to and from the main train station and the city center.

Fig. 12.8: Example of the thermal input image and resulting binary image from one camera view.

From the analysis it is now possible to estimate trajectories of persons on the square. Figure 12.8 shows the results of such trajectories in (a) for one minute, (b) two minutes, and in (c) the velocities vectors for persons walking on the square for a period of one hour. This information may also be illustrated by how the square is occupied, that is not taking the motion of the persons

into account. Such an occupancy map with a resolution of 1×1 meter for 24 hour is illustrated in figure 12.9 sampled for every 10th second.

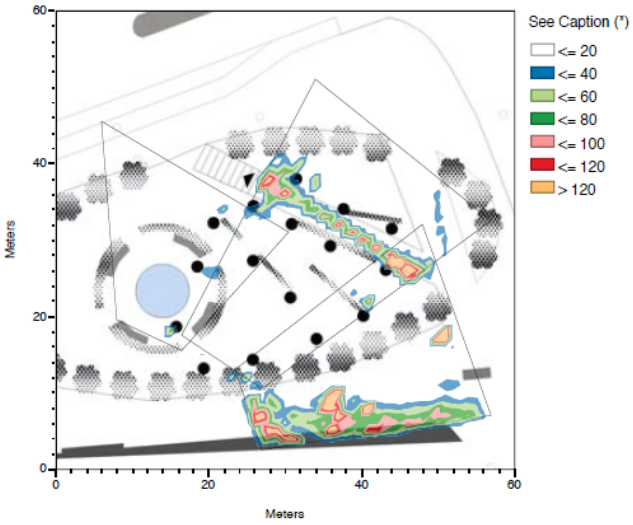


Fig. 12.9: Occupancy map of the square counting number of observed persons in 1×1 meter cells sampled per 10 second.

In this way it is possible to both get information about the instant motion of people on the square and accumulated motion over time. The first information may be used for instant control of the illumination whereas the later representation can be used for placement of lamps, that is the physical design of the illumination setup for a given urban environment, and to design basic illumination according to how the area is being used.

12.2.4 Interactive illumination design

To approach a responsive light design of an urban square calls for a creative process similar to that needed in the development of architectural space. In addition, we need to develop tools to provide creative techniques where interactive scenarios can be sketched and evaluated in a creative and intuitive design process. To approach the design challenge, a physical 1:50 model of the square was developed. Simple white LEDs were used to represent the lamps, and, by using video input recorded on-site by the thermal cameras, we could test how different illumination designs would unfold on the square. In this way, we could simulate the light design using reallife video feeds and evaluate response times, rhythms and the placement of the lamps. When designing interactive illumination for a square, there are multiple factors affecting the formalized lighting, and the illumination becomes a result of many variables including security, social space, functionality, aesthetics, and energy. When these are

merged together in a layered model, one can develop more or less interactive or playful light strategies, which still fulfill functional and aesthetic requirements.

12.2.5 Interactive lighting strategy

To ensure a persistent minimum illumination of the square when there were no occupants, we divided the illumination design into an ambient contribution and an effect contribution, which are later summarized into the final illumination.

Ambient Illumination

The ambient illumination is used to ensure that the square is illuminated when there are no occupants. We followed the hypothesis that a minimum of light is necessary to ensure that people feel that it is secure to enter the square. In the experiment, we worked with two ambient light scenarios:

1. A global minimum; all lamps are dimmed down equally to 10% of the full intensity.
2. Ember; the light slowly fades up and down between 0 to 20% in a random pattern.

Effect illumination

Effect illumination is the response that occur if an event takes place at the square. The event is detected by the computer vision analysis that in turn controls what light response to give according to the activity on the square. One can design a range of different complicated, banal or playful scenarios depending on the level of occupancy, velocity, climate, time of day etc. In this initial experiment, we tested the following two effects:

1. Light circle; as an illuminated aura around the occupants the localized light would secure an illuminated circle on min. 10 meter in diameter. This would allow the occupant to perceive variations in pavement and the face of people passing by, which in turn facilitate a secure navigation and travel over the square.
2. Light wave; As a playful illumination scenario we designed a treasure hunt scenario where two of the lights on the square indicate (blue light) the position of a trigger causing a wave of white light to travel over the square. After 10 seconds, a new blue light will emerge in another location. The hypothesis was to make a playful illumination that engaged people in playful and creative situations.

Final Illumination

Summarizing the intensities from the ambient and effect illumination gives the light emitted from each lamp. If the sum of the two exceeds its maximum, the effect is truncated to 100%. The ambient illumination is active when no one is occupying the actual space. This, however, does not mean that the square is not experienced. Typically, it will be observed from a distance, a balcony, a living room, cafe etc. The illumination can then, with very low power consumption, make light patterns that are embracing, inspiring, scary, natural or just neutral depending on the design intentions. However, when people enter the space effect lighting strategies will secure a suitable illumination that potentially address security, aesthetics and social requirements.

12.3 Experiment

The experiment was conducted in the last week of January 2012 from late afternoon and into the early evening. At this time the sun set at 17.00 and thus collection of data and observations took place from 17.00 and until 20.00. The weather during this period was very cold by Danish standards, ranging from -5 to -10 degrees Celsius and very windy. Data was collected by observations and by logging of the results of the computer vision analysis. Between 100 to 150 persons passed the square each day in the observations period. The observations of them was done from the train and bus station and a bar facing the square. Each day 10 to 15 persons were interviewed.

12.4 Result

During the experiment the experience and effect of the four different light scenarios were investigated. The normal illumination of the square was turned off during the experimental period. See appendix A for video examples.

12.4.1 Scenario 1, "Ambient Illumination"

The first scenario is a homogeneous illumination of the square. The 16 lamps had a static intensity of 80% white light and no effect was added. This light scenario was similar to, what could have been a traditional static illumination of the square, and was motivated by a need to compare the change of flow and social behavior on the square in different light scenarios. It seemed that people did not see the changed illumination and acted like nothing was changed from the everyday life. When people were asked about the illumination only a few recognized the changed lighting, even though the lighting before were very limited.

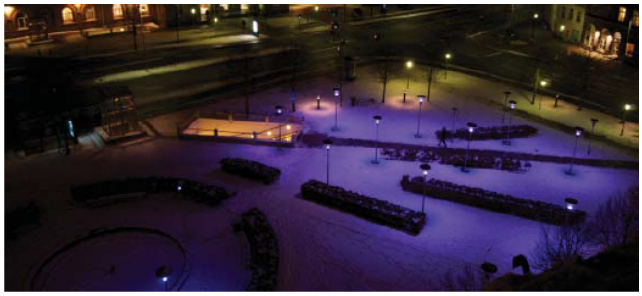


Fig. 12.10: Illumination scenario 1 "Ambient Illumination".

12.4.2 Scenario 2, "Glowing Light"

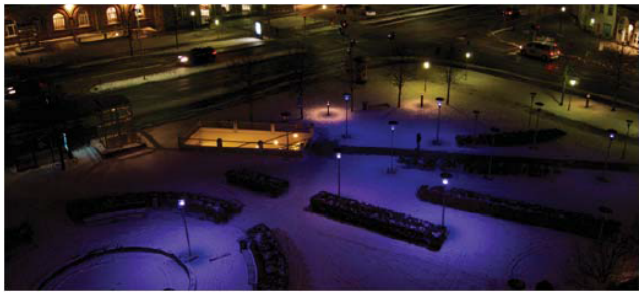


Fig. 12.11: Illumination scenario 2 "Glowing Light".

The intention of the slowly fading white illumination was to make a lighting that would illuminate the square in an aperiodic way, leaving the square half lit but always in a process of fading down or up, this would give a feeling of overview of the square and support the feeling of security (figure 12.11). As a playful chance encounter a light wave effect was introduced. A few people, realized the changing lighting and stopped to look at it, some looked like they thought the light fixtures were broken or out of order, one even asked if there was a loose connection and laughed. The majority of the by-passers did not seem to notice the change in the light intensity when they triggered the light wave; it was like the contrast between the slowly fading lamps in the ember scenario and the slowly moving wave was too small. Observed from a distance the slowly fading lamps had a calming, inspiring and lively expression, one should look very carefully to notice the wave.

12.4.3 Scenario 3, "White Aura"

Because of the relative big illuminated area around the people (10 meters) they did not seem to realize the darker square surrounding them. Observed from a distance one could se how people on the edge of the square were making pointing

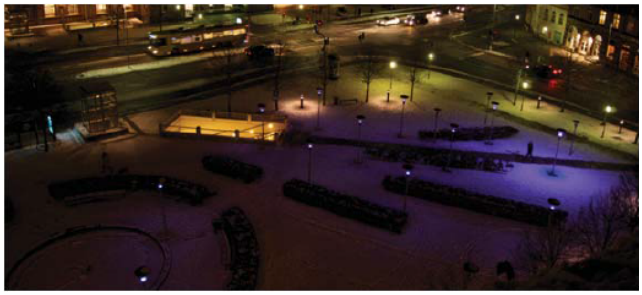


Fig. 12.12: Illumination scenario 3 "White Aura". Notice how the light is following the person on the pathway.

gestures towards the "performing" people moving over the square. The simple effect and the large contrast to the surrounding darkness made the pedestrian a natural focal point on the square.

12.4.4 Scenario 4, "Red Treasure Hunt"

Fig. 12.13: Illumination scenario 4 "Red Treasure Hunt" seen from the perspective of a person walking towards the subway entrance.

The hypothesis of this scenario was to establish an unusual illumination, an illumination that made people stop up and confront the lighting in a playful manner. Figure 12.13 shows a sequence of images visualizing the effect scenario. 1 - A person is approaching the blue "trigger" light. 2 - The person triggered the effect which sends out a wave of light from the triggered lamp outwards. 3,4 - The light wave travels through the square. 5 - The light wave has ended and the trigger point is disabled for 10 seconds before it lights up a lamp again and a new person can trigger yet another light wave. The few people who realized the changing illumination, did engage in the investigation of the lighting. When

they realized that the illumination changed based on their presence, they began calling the lamp "it", which shows that they are giving the lamp a personality.

12.4.5 Energy consumption

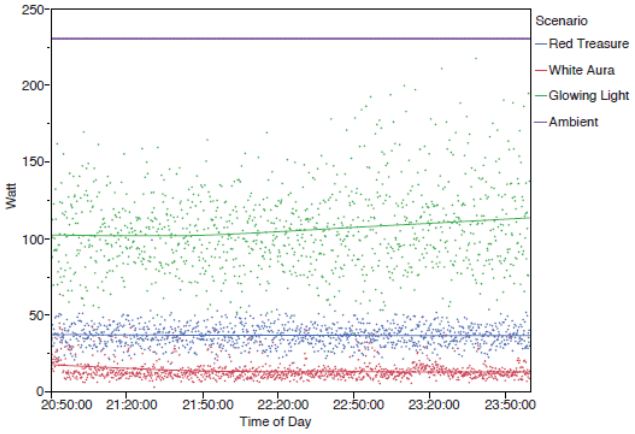


Fig. 12.14: Energy consumption of the four illumination scenarios. The fluctuations around the mean is due to the light effect being set or not set according to the activity at the square and the effect of the changing ambient illumination.

In figure 12.14, the energy consumption for the different light scenarios is illustrated. The Ambient Illumination scenario as reference gave an energy consumption of approximately 230 Watt for the sixteen lamps. The other three scenarios are fluctuating in energy consumption due to the deliberate effects and the reactive effects according to the activity at the square. Clearly, the energy consumption of these scenarios are depending on the light design, and especially the choice of ambient light has a significant contribution of the mean value that the scenario fluctuate around. The order of the energy consumption for the scenarios were that the Glowing Light had a consumption of around 100Watt and this was the scenario with the largest fluctuation that in some cases almost reached the Ambient Illumination level. In the design a lamp is chosen at random and then the light level is set to 10% of the maximum level from which it is then decreased. The Red Treasure Hunt had the second lowest energy consumption and was the scenario that fluctuated most due to the activity at the square whereas the White Aura had the overall lowest energy consumption about 20 Watt in average or 11.5% of the Ambient Illumination scenario. This is also the scenario that mainly depends on the activity at the square.

12.5 Discussion and Future Work

The flow maps of the 4 different light scenarios present no change in movement patterns; there is high intensity of movement on the paths in south and towards the subway, which correlate with the observations, done from the balcony. Hence, we can conclude that the reactive light patterns do not affect the overall movement and use of the square. People did not engage in, what we designed as "playful" and did not engage in the exploration of the reactive light system. This can be due to a less successful design, the fact that people mind their own business, because of the cold evening hours of the Danish winter climate or because of the evening time, that produced a dominate goal oriented behavior. The effect illuminations were directly triggered by the action of individuals, in the initial hypothesis we assumed that these effects would change the behavior of people and affect their way of seeing each other. In the observations of the life on the street, we did observe that people crossing the square, did not realize the changing lighting, but people observing from a distance could see how people "painted" light paths on the square or triggered a wave. The observer could be a person in a bus, on a balcony, in the waiting room for the cinema etc. For them the square became a stage setting where actors, whiteout being aware of it affected the lighting and was observed more carefully. This observation supports the concept of passive surveillance as presented by Jane Jacobs [28]. These observations leave the light designers with a new challenge: How can we design adaptive systems that change behavior due to changing situations and needs? In times of low occupancy a simple and functional reactive light strategy might be sufficient whereas in a warmer climate interactive playful scenarios might be applicable. Furthermore, at public events the street lighting might be used as an extension of the stage lighting. Extending the notion of centralized light control in to the world of smart phones, one can imagine a light system that knows your personal preferences for lighting and prepare the city or park with your favorite settings, as such it becomes an extension of your expression of mood, identity or just favorite color. These possibilities are subjects for future experiments. Furthermore the experiment underline the hypothesis that reactive and sensor based lighting systems will save energy and the amount of energy saved depends on the designed light scenarios. In the experiment we present an energy saving of 92%. However to access this resource new robust technologies need to be developed. This contribution encourages a stronger link between studies in the contemporary interactive art scene and the behavior design of public lighting. The authors have not been able to find systematic research in the field of experienced social and aesthetic qualities of responsive light design in the "smart city" context. When we still have the centralized control unit for sketching light scenarios, a wide range of design drivers has the capacity to control the illumination and there are three main challenges in the future work: 1) to further develop design scenarios, and develop the ethnographical and qualitative evaluation techniques of such, 2) to face the technical challenge and build simple reactive and robust

stand alone solutions, which perform sensing and acting behaviors, and 3) we suggest further development and explorations of the flow maps, which can inform us about usage of space instantly and over time and thermal comfort. To approach these design challenges will call for interdisciplinary research, where engineers and architects will work together for development of new robust tools and novel design methods, to be tested in the 'laboratory of the street'. Favoring a design side of mediated lighting can cause inefficient light tools and favoring a technical side would risk to create normative and boring environments. Embarking in this interdisciplinary journey, we will search for skills to develop new tools and techniques to modulate a new creative environment.

12.6 Conclusion

Street lighting is build to illuminate the square extending the potential use of public space into the dark hours. Until now we have focused on the energy performance of the light bulb street lamp. Because of the recent development in the field of sensor and LED technologies we are now able to modulate the light to any given control paradigm. This study shows new possibilities and reflections for applying simple reactive light strategies, together with pragmatic reactive lighting models in the design of interactive urban lighting scenarios. It does this through four experiments using thermal cameras and computer analysis that allow designers to detect occupancy and flow patterns on the street. The data is utilized both as input to a real-time light control system and as a mapping of long-term occupancy and flow, allowing researchers and urban planners to access data on the use of urban spaces. In this paper, the evaluation of three interactive light strategies; Glowing Aura, Glowing Light & Red Treasure Hunt, reveal the possibility for reactive lighting to be applied in public spaces and present a significant energy saving up to 90% without people changing behavior. This result shows that very dramatic changes are needed if light designers are to engage a person in transit. However people making light patterns as actors on a stage, underline the points of Presentation of Self in Everyday Life by Erving Goffman [29]. The majority of the visitors did not realize the changing of the light, the first time of their visit, but after observing other people perform from a distance their occupancy and flow patterns became a natural part of the architectural expression of the square.

Acknowledgments

This work was funded by Aalborg University and Center of Danish Lighting in collaboration with private collaborators Team Tronic, Riegnes, Alfred Priess, Thanks to Nykredit and The municipality of Aalborg for practical support. The experiments were carried out at Research Cluster for Mobility and Tracking Technologies (MOTT), Aalborg University.

Appendix

For video examples of the scenarios explored in the experiment please go to the following urls:

Glowing Light

http://vimeo.com/40589332

White Aura

http://vimeo.com/40653666

Red Treasure Hunt

http://vimeo.com/40589333

References

[1] O. B. Jensen, "Flows of meaning, cultures of movements - urban mobility as meaningful everyday life practice," *Mobilities*, vol. 4, no. 1, pp. 139–158, 2009.

[2] ——, "Negotiation in motion: Unpacking a geography of mobility," *Space and Culture*, vol. 13, no. 4, pp. 389–402, 2010.

[3] N. Wiener, *Cybernetics*. New York: Wiley, 1948.

[4] R. C. Arkin, *Behaviour-based Robotics*. London: MIT press, 1998.

[5] N. Negroponte, *Architecture Machine*. Cambridge: MIT Press, 1970.

[6] A. Greenfield and M. Shepard, "Urban computing and its discontents," *Situated Technologies Pamphlets 1*, 2007.

[7] G. Pask, *An approach to cybernetics*. New York: Harper, 1961.

[8] U. Haque, *4dsocial: Interactive Design Environments*. Wiley, 2007.

[9] P. Pangaro, "Thoughtsticker," *Kybernetes*, vol. 30, no. 5/6, pp. 790–807, 2001.

[10] C. Linn, "MY studio/Höweler + Yoon Architecture's Lo Rez Hi Fi gives a D.C. building a much-needed sense of place." *Architectural Record*, vol. 195, no. 11, pp. 190–192, 2007.

[11] B. S. Thomsen, *Performative Environments*. PhD Thesis, Doctoral School of Planning and Development, Aalborg University, 2009.

[12] P. Beesley, *Kinetic Architectures and Geotextile Installations*, 1st ed. Toronto: Riverside Architectural Press, 2010.

[13] A. Chong and T. de Rijk, *Daan Roosegaarde: Interactive Landscapes*, 1st ed. Rotterdam: NAI Publishers, 2011.

[14] Echelon. (2007) Monitored outdoor lighting: Market, challenges, solutions, and next steps. [Online]. Available: http://info.echelon.com/Whitepaper-Monitored-Outdoor-Lighting.html

[15] IBM, *Smarter Planet*, 2012.

[16] European Commission, *European SmartGrids Technology Platform - Vision and Strategy for Europe's Electricity Networks of the Future.* European Commission, 2006.

[17] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Fourth IEEE International Conference on Automatic Face and Gesture Recognition. Proceedings.*, 2000, pp. 46–53.

[18] Z. Niu and X. Qiu, "Facial expression recognition based on weighted principal component analysis and support vector machines," in *3rd International Conference on Advanced Computer Theory and Engineering (ICACTE)*, vol. 3, Aug 2010, pp. V3–174–V3–178.

[19] W. Wang, J. Zhang, and C. Shen, "Improved human detection and classification in thermal images," in *17th IEEE International Conference on Image Processing (ICIP)*, Sept 2010, pp. 2313–2316.

[20] E. Goffman, *Behavior in public places : notes on the social organization of gatherings.* Westport: Grenwood Press, 1980.

[21] E. T. Hall, *The Silent Language.* New York: Anchor Press, 1973.

[22] J. Gehl, *Cities for People.* Island Press, 2010.

[23] W. H. Whyte, *City: Rediscovering the Center.* Philadelphia: University of Pennsylvania Press, 1988.

[24] M. Piccardi, "Background subtraction techniques: a review," in *IEEE International Conference on Systems, Man and Cybernetics*, vol. 4, Oct 2004, pp. 3099–3104.

[25] A. Criminisi, "Computing the plane to plane homography," 1997. [Online]. Available: http://www.robots.ox.ac.uk/~vgg/presentations/bmvc97/criminispaper/node3.html

[26] J. A. Hartigan, *Clustering Algorithms.* Wiley, 1975.

[27] G. Welch and G. Bishop, "An introduction to the kalman filter," Chapel Hill, NC, USA, Tech. Rep., 1995.

[28] J. Jacobs, *The death and life of great American cities.* New York: Random House, 1961.

[29] E. Goffman, *Presentation of self in everyday life.* Garden City, N.Y., Doubleday, 1959.

Chapter 13

Taking the Temperature of Pedestrian Movement in Public Spaces

Søren Zebitz Nielsen, Rikke Gade, Thomas B. Moeslund and Hans Skov-Petersen

The paper has been published in
Transportation Research Procedia, Vol. 2, pp. 660–668, October 2014.

© 2014 Elsevier
The layout has been revised.

Abstract

Cities require data on pedestrian movement to evaluate the use of public spaces. We propose a system using thermal cameras and Computer Vision (CV) combined with Geographical Information Systems (GIS) to track and assess pedestrian dynamics and behaviors in urban plazas. Thermal cameras operate independent of light and the technique is non-intrusive and preserves privacy. The approach extends the analysis to the GIS domain by capturing georeferenced tracks. We present a pilot study conducted in Copenhagen in 2013. The tracks retrieved by CV are compared to manually annotated ground truth tracks, and an example of pedestrian behavior is analyzed.

13.1 Introduction

Planners designing well-functioning liveable cities for people need to know how streets and public spaces are being used and how pedestrians move. The classic approach to collect such data is to make sample counts of people at points of interest a few times a year and conduct qualitative urban analysis [1, 2]. With the rapid development of computing and networking technologies, the miniaturization of sensors, and the introduction of smartphones, a range of new ways to capture data on people's movement have become available in recent years with potential to supplement and extent the classic methods.

Several studies that track people by using data from smartphones and their signals and sensors, such as Bluetooth, Wi-Fi and Global Navigation Satellite System (GNSS), have been made [3–7]. These studies are interesting on a city wide scale to understand macro movement patterns of samples of people, but the spatial accuracy of data from smartphones is not good enough to study detailed pedestrian movement patterns and behaviors in urban streets and plazas. This instead requires accurate and simultaneous tracking of several individuals who may move close together, and where the movement of each individual depends upon interactions with others as well as on the physical layout of the place and attractors in the space traversed [8, 9]. For such micro scale pedestrian studies Computer Vision (CV) based tracking technology is more appropriate to use as it is able to passively register the activity in a plaza without affecting the behavior of people in the space. However there are several challenges to capture reliable data this way. This paper presents results of capturing and exploring data on pedestrian movement with CV tracking technology from a pilot study we conducted in the summer of 2013 in the urban plaza 'Kultorvet' in central Copenhagen (see Fig. 13.1).

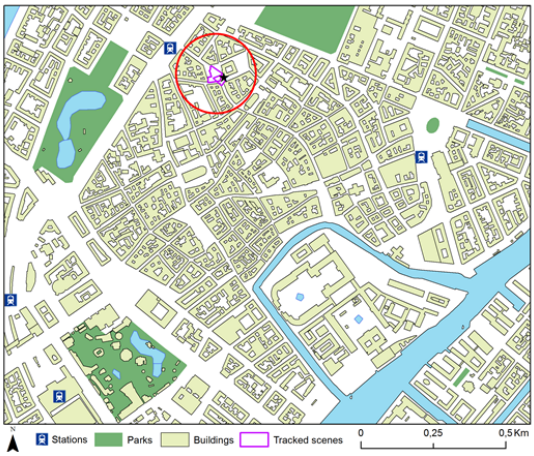


Fig. 13.1: Overview map of central Copenhagen with the tracked area at 'Kultorvet' highlighted. The background data for this map and the buildings in Fig. 13.5, 13.6, and 13.7 are taken from the open public geographic data courtesy of the Danish Geodata Agency.

13.2 Methods

As Computer Vision technology has made rapid progress in recent years [10–12] we wanted to test it in studies of pedestrian movement patterns and behaviors in everyday traffic in public spaces to assess its potential as a tool to capture such data and aid planners in future Smart Cities [13]. Camera surveillance of public spaces in the form of CCTV systems is already installed in many cities, but these systems are often calibrated to aid the police in crime fighting or as traffic cameras to identify vehicles and report on the traffic situation. Police cameras are often Pan-Tilt-Zoom (PTZ) cameras which can be used to zoom in on situations and identify suspects and follow them within a network of cameras around the city, and traffic cameras are focused on vehicle traffic on the road network. To conduct pedestrian movement studies with CV tracking it is necessary to have dedicated cameras with a fixed Field of View (FOV) that can be set up to constantly monitor an Area of Interest (AOI), such as an urban plaza, preferably from an elevated position. This in order to get close to a nadir looking position to avoid the occlusions that occur when people pass each other in front of the camera as CV algorithms can have difficulties to handle occlusions and to distinguish individuals that move close together. The height for optimal camera installation is also a balance between how large a FOV needs to be surveyed versus the level of detail that can be seen in the image. The fact that the further away objects are in a camera's FOV the smaller they appear in the image needs to be taken into account since it is more difficult for CV algorithms to detect and distinguish individuals if they only take up few pixels in the image [14].

Computer Vision algorithms can be applied on video from both normal RGB cameras and thermal cameras. In terms of performance of CV algorithms there are advantages and disadvantages in both technologies that need to be considered [15, 16]. Normal RGB cameras record reflected light in three channels, one for the red, green, and blue color respectively, and therefore these cameras depend on sufficient light to operate. RGB sensors are cheap, but they are also somewhat complicated to use for CV tracking of humans in urban scenes. Light conditions change between day and night, and they can change fast in different weather situations or when the scene is illuminated from headlights on a moving vehicle etc. This is a challenge for CV algorithms that depend on a reliable background model to segment moving objects from the background by the use of background subtraction. Segmentation using background subtraction assumes that it is possible to obtain a reliable background model and that only people, or other objects of interest, are moving. A dynamic background model can be adjusted and updated during run-time of the algorithm, but it is often not enough to achieve good results with RGB cameras. Another well-known problem for CV in the RGB-domain is the occurrence of shadows which often cause false detections, as the shadows move just like people.

In the thermal domain there are no shadows and no issues with fast changing lighting conditions as thermal cameras record the long-wave infrared radiation (8-15μm) emitted by all objects with a temperature above absolute zero. Thermal cameras can thus detect people and objects with a temperature different from the surroundings both day and night independent on the light in the scene. The concept of establishing a background model and use background subtraction to segment moving objects is the same in the thermal domain, and it is easier here to model the background and update it dynamically as temperature conditions in a scene often change more slowly compared to lighting conditions. Dynamic updating of the background model is necessary in outdoor scenes as the surrounding temperature change throughout the day as well as the sun can heat dark pavements to temperatures hotter than the human body temperature. Therefore it cannot just be assumed that people are constantly hotter or colder than the background and thus the CV algorithm has to adjust accordingly. Still people can often more reliably be segmented with background subtraction using thermal cameras in comparison with RGB cameras, and the method is computationally fast, making it well suited for real-time applications. In terms of false detections in thermal images there can be issues for glossy surfaces as thermal radiation can be reflected from these. This is however considered a rare problem in Smart Cities applications as the surfaces are often not reflective and those that are, such as windows, are permanent installations so data from these areas in a scene can easily be filtered out.

When it comes to resolving occlusions and ambiguous situations the RGB cameras have an advantage over the thermal cameras. The three channel RGB images contain more information compared to the one channel of thermal cameras. Colors and textures extracted from RGB images can be applied to re-identify individuals after occlusions, which is not possible for thermal images.

However, tracking can often be simplified to the detection of an object, and then assigning the detection to a path established in previous frames by considering the velocity and direction of the object. Where perfect re-identification of objects is not required tracking in thermal imaging can perform well. At the same time the inability to identify individuals in thermal imaging is also one of its greatest assets because privacy issues related to recording video in public places can be neglected as there is no risk of revealing individuals identity in the thermal images. Privacy by design is thus ensured when using thermal cameras for tracking. The pros and cons of RGB versus thermal cameras are summarized in table 13.1 below.

	Pros	**Cons**
RGB	Cheap sensors	Sensitive to light
	Re-identification possible	Privacy issues
		Shadows
Thermal	Easier segmentation	Re-ident. difficult
	Independent of light	More expensive
	No privacy issues	Reflections
	Single channel images	

Table 13.1: Overview of pros and cons of RGB versus thermal camera for computer vision tracking applications in urban outdoor scenes.

Even though thermal cameras are more expensive than RGB cameras, especially because of the germanium metalloid needed for the lenses, they have some clear advantages in CV tracking applications [16]. With the development in thermal camera technology the resolution of thermal cameras is slowly increasing as the technology evolves and new materials are explored. The cost is also lowering and will probably continue to do so and make them more cost effective to deploy. Considering the capabilities of thermal cameras we wanted to explore their potentials and pitfalls, as well as the quality of the data they create, in a real life CV tracking experiment on pedestrian movement patterns and behavior in urban public spaces. Related studies on pedestrian tracking using thermal cameras are [17] and [18]. However these studies use slightly different Computer Vision techniques and are applied in a different context.

In our pilot study we used one state-of-the-art uncooled passive thermal camera with a resolution of 640×480 pixels (Axis Q1922), a lens with a focal length of 10 mm, a viewing angle of 57°, and 30 fps camera frame rate. Background subtraction was applied to detect people. A background model was obtained by calculating the median value for each pixel over a 30 second initializing period. The background was updated during run-time, using a selective update method, meaning that only pixels segmented as background contribute to the updated background. The foreground objects were filtered by size, in order to remove noise. In order to solve partial occlusions, we were able to split BLOBs both vertically and horizontally [19]. Binary Large Objects (BLOB) refers to a group of connected pixels in a binary image [20]. After

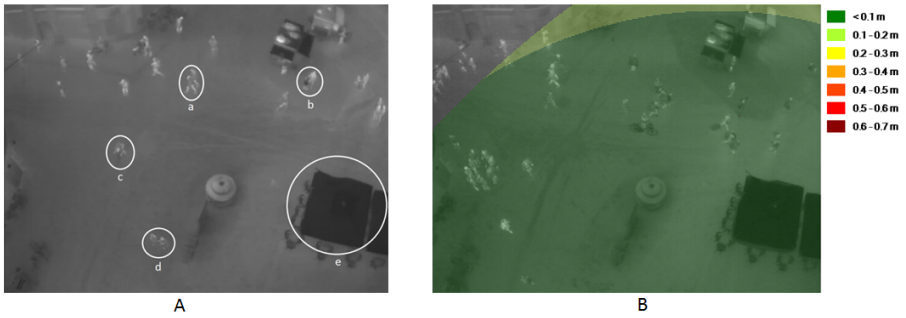


Fig. 13.2: The near nadir view. The areas highlighted in A refer to examples of behaviors and situations. The image in B is generated in T-analyst and shows the uncertainty for the pixels in the image in relation the real world coordinates.

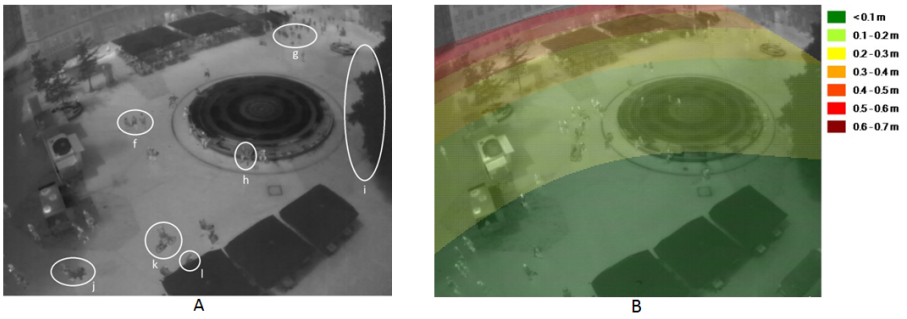


Fig. 13.3: The view overlooking the plaza. The areas highlighted in A refers to examples of behaviors and situations. The image in B is generated in T-analyst and shows the uncertainty for the pixels in the image in relation the real world coordinates.

converting the position of the remaining objects to world coordinates, they were tracked using Kalman filtering [21]. The processing time of the computer vision tracking algorithm was 20ms per frame on an Intel Core i7-3770K 3.5 GHz CPU with 8GB RAM. This is fast enough to provide the raw tracking data in real-time, even though this capability was not tested in our pilot study.

A coordinate system is needed in order to relate the captured tracks to each other and the surrounding space. In a fixed FOV the pixel coordinates in the camera image will remain pointed at the same locations in the scene. By measuring control points in the scene homography can be used to relate the pixel coordinates in the image plane to the real world coordinates in the ground plane by use of a transformation matrix [22]. To georeference the scene we measured the control points with high precision GPS equipment. With a georeferenced scene the movement data can be related to the surroundings and other geospatial data layers in Geographic Information Systems (GIS), which also allow for spatio-temporal analysis of the data. Depending on people's position in the FOV the spatial accuracy for each pixel is between 0.1 and 1

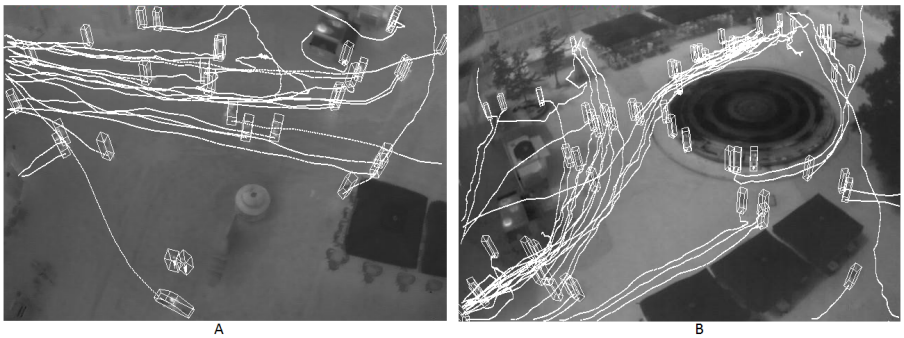


Fig. 13.4: The two views with the GT tracks digitized manually in T-Analyst overlaid.

meter (see Fig. 13.2B and Fig. 13.3B).

For each frame, the CV tracking software yields a list of ID numbers and positions of the detected persons in real world coordinates. To read the data in a GIS, the raw text files from the CV software were passed to a Python script to render a list of locations for the tracking of each of the IDs. To reduce the amount of data processed while still maintaining sufficient location accuracy the points were down-sampled to five track points per second than the original rate of 30 points per second. During parsing of the raw files, a series of attributes were added to the individual points, including speed (in relation to the previous point, a given interval back in time and accumulated for the track up to the given point) and incremental distance and time. Further metadata were generated for each individual track, including distance, duration, Euclidean distance, average speed, number of points etc.

To assess the quality of the CV trajectories in terms of their completeness and accuracy the CV tracks needed to be able to be evaluated against Ground Truth (GT) trajectories. To do so we have manually digitized the GT trajectories of all individuals in the video recordings presented in this study. This has been done in T-Analyst developed at Lund University [23]. In this software pedestrians are modeled as 3D rectangles with the dimension of 0.5x0.5x1.8 meters (see Fig. 13.4). The user can then manually digitize the position of a pedestrian frame by frame. The GT tracks digitized in T-Analyst were also transformed from pixel to real world coordinates in a similar manner as for the CV tracks and imported into GIS.

13.3 Scene Description

The plaza 'Kultorvet' was used as a test scene. It is in a pedestrian zone in central Copenhagen with occasional bicycle traffic and goods delivery by vehicles. The part of the scene closest to the camera was situated where one of the city's major shopping streets meets a perpendicular street at the entrance

to the open plaza. The street at the far end of the plaza leads directly to the subway station with the most traffic in the city. The scene had a continuous flow of pedestrians (50-100 per minute) coming from several directions that needed to negotiate and avoid each. The thermal camera was placed on the roof top terrace of a five story building. Two views were recorded. The first one, which we label the near-nadir view, was a straight down view to get as close to the nadir position as possible in order to minimize the number of people occluding each other in the camera FOV (Fig. 13.2A). Consecutively the second view, labeled the overlooking view, was taken from the same spot but now instead overlooking the entire plaza from an oblique angle (Fig. 13.3A). The scenes were recorded around noon on a Friday. The weather was overcast with occasional showers of rain.

Tracks of people walking alone or in social groups of different sizes were recorded (Fig. 13.2A a and Fig. 13.3A f), as well as people sitting or waiting (Fig. 13.3A h), people having a conversation (Fig. 13.2A d and b), and people dragging their bikes (Fig. 13.2A k) or pushing a pram or stroller (Fig. 13.3A j). The tracks of 'facers' working for a charity organization trying to stop people in the street to make them donate to the cause were also recorded in the scene (Fig. 13.2A b). Occasionally cyclists riding through the scene despite the legislation were observed (Fig. 13.2A c).

Following the video recordings the permanent objects in the scene such as hotdog stands, sun shades, benches, a fountain etc. were digitized as polygons in GIS from the most recent orthophotos of the place and cross checked with the thermal video recordings to confirm their actual positions (grey objects in Fig. 13.5, Fig. 13.6, and Fig. 13.7). In this way these objects could be drawn and georeferenced in GIS and the tracks thus analyzed in relation these objects. A layer with the surrounding building footprints were added and drawn as well (green polygons in Fig. 13.5, Fig. 13.6 and Fig. 13.7).

13.4 Analysis and Results

Half an hour of thermal video was recorded for both views. It was decided to only process five minutes of the video from the first view and one minute from the second view. The reasons for processing only these sections was mainly due to long time it took to manually digitize the GT tracks for assessment of the CV tracks. However it was considered that this was enough to get data on the general movement patterns as well as to identify areas and standard situations in the scenes in which the CV software was challenged.

The data generated for the CV and GT tracks were imported to a geodatabase in the ArcGIS software as a point feature class. For each point record the basic data was in the format (*FrameNumber, ID, Xcoordinate, Ycoordinate, TimeStamp*). Furthermore the derived variables in terms of the accumulated time, distance and average speed for the track were written for each point record. The Python script passing the data also generated a table summariz-

ing statistics such as total distance, total duration, average speed, start and end frame number, and number of track points for each *ID*. The track point table and the statistics table were joined on the *ID*. To enable visualization of the temporal dimension the data was converted to 3D by using the *FrameNumber* attribute as the Z coordinate. Since the geodatabase date-time format did not natively support milliseconds the *TimeStamp* field could not be used. To connect track points to lines a standard point to line tool based on the *ID* and sorted by the *FrameNumber* was applied.

Inspired by the works on visual analytics of movement by [24] and [25] the data was explored visually in 2D and 3D. The tracks obtained for the near nadir view and the overlooking view are shown in Fig. 13.5 and Fig. 13.6 respectively. The A parts in the figures depict the tracks obtained from the CV algorithm in a 2D map, the B parts show the manually digitized GT tracks for the same scene, and the C parts show the data from A and B plotted in a Space-Time Cube (STC) for visualization of the temporal dimension. For Fig. 13.5C only the first minute of data is shown in the STC, but data for all five minutes are shown in Fig. 13.5A and Fig. 13.5B. A visual exploration of the data is naturally best undertaken in a 3D GIS program where one can work with the STC to rotate, pan and zoom in on interesting areas of the cube. It is much more challenging to communicate the results visually on paper. The following describes our visual inspection and interpretation of the data to identify areas and situations in the scenes where the CV software had difficulties in tracking individuals correctly.

To filter out noise in the CV data it was decided to accept only CV tracks with duration of three or more seconds. This approach identified most of the ambiguous situations in which people had occluded each other from view which confused the CV software. The situations occurred most often in the area with the main flow of people i.e. the highest density of tracks as seen on Fig. 13.5A, Fig. 13.5B and Fig. 13.6A, Fig. 13.6B. The short detections are plotted as red dots. A prominent red area is spotted near the camera in Fig. 13.5A. This was caused by the two people marked with d on Fig. 13.2A, who stood close on the same spot and talked for all five minutes of video, which gave the CV algorithm a challenge to obtain a lock on them as two individuals. Instead it saw them as part of the background in most frames and assigned them new track IDs each time they moved slightly. The same was the case for the sitting people marked with h in Fig. 13.3A. These two are thus examples of false negative detections.

It was evident from the data that there were areas with a large amount of detections that could not be moving people, but instead movement of objects such as tree branches (area i on Fig. 13.3A) or the canvas on sun shades (areas e and l on Fig. 13.2A and 13.3A) swaying in the wind. These were thus all tagged as red short detections as seen in the upper grey area depicting the sun shades in Fig. 13.5A, and in the long dense red area in Fig. 13.6A that is due to the tree branches. These areas are thus examples of false positive detections.

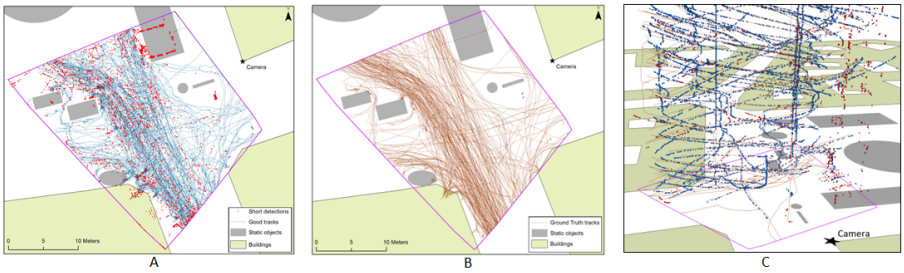


Fig. 13.5: The tracks from the near nadir view. A shows the good CV tracks as well as the short CV detections. B shows the manually digitized GT tracks for comparison to A. C shows a Space-Time Cube of the first minute of data from A and B.

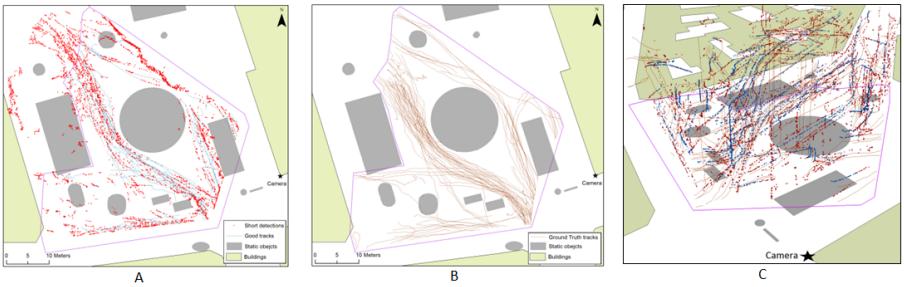


Fig. 13.6: The tracks from the view overlooking the plaza. A shows the good CV tracks as well as the short CV detections. B shows the manually digitized GT tracks for comparison to A. C shows a Space-Time Cube of the data from A and B.

Since it was difficult for the CV algorithm as well as manually to track individuals in the far end of the plaza in the overlooking view (area g in Fig. 13.3A), it was decided to draw a polygon in which at least unambiguous manual GT tracks could be digitized (see Fig. 13.6). All CV detections outside this area were tagged as short detections and clipped from the good CV tracks inside the polygon.

In our manually digitized GT we found 297 tracks crossing the near nadir scene in the 5 minutes of data. The CV algorithm found 460 tracks that could be classified as good tracks. 1475 track IDs were classified as short detections. Close inspection of the STC showed that several of the good CV tracks could be referred to as being parts of one corresponding GT track. This indicates the IDs of good CV tracks have been changed after ambiguous situations, but that the same individuals have been tracked all through the scene. In the one minute data from the overlooking view we found 124 tracks by the manual GT method and the CV method found 146 tracks. There were 977 IDs classified as short detections.

The overall movement patterns detected by the CV fit well with the GT tracks, and in the STC in Fig. 13.5C and Fig. 13.6C it can be seen that

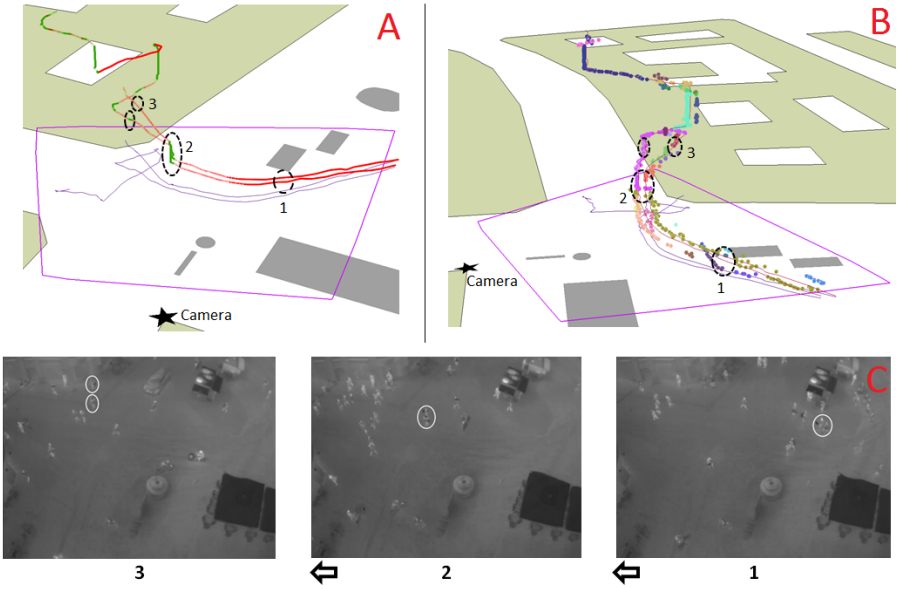


Fig. 13.7: Example of extraction of the two tracks of a couple. The GT tracks are shown in A, the CV track points are shown along with the GT tracks in B, and the thermal images for three step sequence is shown in C.

individual GT and CV tracks are fairly good aligned. A quick inspection (not shown) of the speeds of the good CV tracks yielded a Gaussian distribution around 5km/h for pedestrian tracks. A small spike around 13 km/h was clearly identified as cyclists in the scene.

The data also allowed for extraction of individual tracks for analysis of movement behaviors seen in the video. An example of this is shown in Fig. 13.7 where the two tracks of a couple are shown and where three situations from their tracks are highlighted. In situation 1 they walk together, in 2 they stop to discuss and one points at the fashion shop on the street corner, in 3 the two tracks split up as the woman (verified by live observation when the scene was recorded) decides to enter the shop while the man waits outside. The thermal images of the situations are shown in Fig. 13.7C. In Fig. 13.7A the GT tracks are shown in a 3D plot with the two tracks colored in red when they move more than 2 km/h and green when they move slow or stand still. The tracks are projected on the 2D plane for reference. In Fig. 13.7B the CV track points detected within a radius of one meter and one second, in the spatial and temporal dimensions respectively, of the corresponding GT tracks are displayed. The different CV tracks are colored according to their IDs. It is evident that the IDs are changed quite a few times for the CV track points, even though the two persons remained tracked throughout the sequence. The CV method is thus not robust enough yet for tracking individuals consistently

in advanced and ambiguous situations, but it performs fairly well to extract overall movement patterns.

13.5 Discussion and Conclusion

In this paper we have presented a method using thermal cameras and state-of-the-art Computer Vision technology to track pedestrians in public plazas. The tracks are georeferenced to enable analysis in GIS. Furthermore we have compared the tracks obtained from Computer Vision with manually annotated ground truth tracks by visual analysis of the data on 2D maps and in 3D Space-Time Cubes. The analysis showed examples of situations in which the Computer Visions tracking was challenged, but indicated also that the method has potential to provide reliable data on general movement patterns in a plaza with a similar density of people. An example of extraction and analysis of two tracks of a couple in a standard situation was shown. From this it was clear that the Computer Vision software was not able to keep the individuals tracked consistently, however it was able to detect the persons present all the time. The changing of track IDs was most likely caused by the two people walking and standing close together as well as by others passing close by. It could be useful to test our method in other types of plazas and with higher densities, and to extract and analyze tracks from different situations to compare data.

Further development of the method could be to enable computation of statistics for the relation between Computer Visions and ground truth tracks in terms of accuracy and completeness. An idea could also be to enable automatic classification of moving individuals into types such as pedestrians, cyclist, people pushing prams etc. The dataset obtained in this pilot study has enabled a better understanding of the difficulties for further automation. While going through the videos to digitize ground truth tracks we identified characteristic movement behaviors such as meeting, flocking, avoidance, and following a leader [26]. Further research could hence be in the field of Computational Movement Analysis to automate search for characteristic movement patterns and behaviors in this kind of dataset.

We are working on plans to carry out a full scale study with multiple thermal cameras over a sustained period of time. Inspired by the works of [2] and [27] the ambition is to contribute with new digital methods and tools in the field of urban analysis by linking the field of Computer Vision to that of GIS. The long term goal is to further investigate the applicability and suitability of this type of studies to provide data for models of everyday pedestrian behavior in urban public spaces.

Acknowledgements

We would like to thank Aliaksei Laureshyn from Lund University for permission to use the T-Analyst software and assistance with setting it up. We would also like to thank the administration of the building Knud Højgaards Hus in Copenhagen and the company Fokustranslatørerne for allowing access to their roof top terrace for us to install the thermal camera for the pilot study. Thanks to Louise Schnegell for help with setting up the manuscript in LaTeX.

References

[1] D. Bauer, N. Brändle, S. Seer, M. Ray, and K. Kitazawa, "Measurement of Pedestrain Movements: A Comparative Study on Various Existing Systems," in *Pedestrian Behavior*, H. Timmermans, Ed. Emerald Group Publishing Limited, 2009, ch. 15, pp. 325–344.

[2] J. Gehl and B. Svarre, *How to Study Public Life*. Island Press, 2013.

[3] M. Delafontaine, M. Versichele, T. Neutens, and N. Van de Weghe, "Analysing spatiotemporal sequences in Bluetooth tracking data," *Applied Geography*, vol. 34, pp. 659–668, May 2012.

[4] F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti, "Unveiling the complexity of human mobility by querying and mining massive trajectory data," *The VLDB Journal*, vol. 20, no. 5, pp. 695–719, Jul. 2011.

[5] N. Shoval, "Tracking technologies and urban analysis," *Cities*, vol. 25, no. 1, pp. 21–28, Feb. 2008.

[6] J. van Schaick and S. van der Spek, Eds., *Urbanism on Track*. Amsterdam: IOS Press, 2008.

[7] P. a. Zandbergen, "Accuracy of iPhone Locations: A Comparison of Assisted GPS, WiFi and Cellular Positioning," *Transactions in GIS*, vol. 13, pp. 5–25, Jun. 2009.

[8] M. Moussaïd, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz, "The walking behaviour of pedestrian social groups and its impact on crowd dynamics." *PloS one*, vol. 5, no. 4, p. e10047, Jan. 2010.

[9] H. Timmermans, Ed., *Pedestrian Behavior*. Emerald Group Publishing Limited, 2009.

[10] D. Gowsikhaa, S. Abirami, and R. Baskaran, "Automated human behavior analysis from surveillance videos: a survey," *Artificial Intelligence Review*, Apr. 2012.

[11] T. Ko, "A Survey on Behaviour Analysis in Video Surveillance Applications," in *Video Surveillance*, W. Lin, Ed. InTech, 2011.

[12] T. B. Moeslund, A. Hilton, V. Krüger, and L. Sigal, Eds., *Visual Analysis of Humans: Looking at People*. Springer, 2011.

[13] M. Batty, K. W. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, and Y. Portugali, "Smart cities of the future," *The European Physical Journal Special Topics*, vol. 214, no. 1, pp. 481–518, Dec. 2012.

[14] T. Ko, "A survey on behavior analysis in video surveillance for homeland security applications," *2008 37th IEEE Applied Imagery Pattern Recognition Workshop*, pp. 1–8, 2008.

[15] J. Davis and V. Sharma, "Robust detection of people in thermal imagery," *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, pp. 713–716 Vol.4, 2004.

[16] R. Gade and T. B. Moeslund, "Thermal cameras and applications : a survey," *Machine Vision and Applications*, no. 25, pp. 245–262, 2014.

[17] C. N. Padole and L. a. Alexandre, "Motion Based Particle Filter for Human Tracking with Thermal Imaging," *2010 3rd International Conference on Emerging Trends in Engineering and Technology*, pp. 158–162, Nov. 2010.

[18] P. Skoglar, U. Orguner, D. Törnqvist, and F. Gustafsson, "Pedestrian tracking with an infrared sensor using road network information," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, p. 26, 2012.

[19] R. Gade, A. Jø rgensen, and T. B. Moeslund, "Occupancy analysis of sports arenas using thermal imaging," *Proceedings of the International Conference on Computer Vision and Applications*, 2012.

[20] T. B. Moeslund, *Introduction to Video and Image Processing. Building Real Systems and Applications*. Springer, 2012.

[21] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems 1," vol. 82, no. Series D, pp. 35–45, 1960.

[22] A. Criminisi, "Computing the plane to plane homography," 1997. [Online]. Available: http://www.robots.ox.ac.uk/~vgg/presentations/bmvc97/criminispaper/node3.html

[23] A. Laureshyn, "T-Analyst software," 2013. [Online]. Available: http://www.tft.lth.se/video/co_operation/software/

[24] G. Andrienko, N. Andrienko, P. Bak, D. Keim, and S. Wrobel, *Visual Analytics of Movement*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

[25] D. Keim, G. Andrienko, J.-d. Fekete, G. Carsten, J. Kohlhammer, and G. Melancon, "Visual Analytics : Definition , Process , and Challenges," in *Information Visualization*, A. Kerren, Ed. Berlin, Heidelberg: Springer, 2008, pp. 154–175.

[26] J. Gudmundsson, P. Laube, and T. Wolle, "Computational Movement Analysis," in *Springer Handbook of Geographic Information*, 2012, pp. 423 – 438.

[27] W. H. Whyte, *The Social Life of Small Urban Spaces*. New York: Project for Public Spaces Inc, 1980.

Part VI

Conclusion

Chapter 14

Conclusion

This thesis has presented methods to automatically analyse the human motion and activities captured with thermal cameras. With main focus on the analysis of sports arenas, the work was organised within three themes; Occupancy analysis, Activity recognition and Tracking sports players.

Two methods were presented within the topic of occupancy analysis. While the first method analysed each frame individually and suggested a number of steps for splitting and sorting blobs, the second method included temporal information. A graph-based approach was proposed to optimise the result over a sequence of frames.

In the third part of this thesis, two different methods for recognition of activities were presented. Based on heatmaps, the first method relied only on the position of people. The second method were based on features extracted from tracklets. Five sports types were classified in each work.

In the topic of tracking, this thesis presented two methods designed for thermal video, and a third method improving a global tracking algorithm for RGB data. The first method based on Kalman filtering showed robust real-time performance and high accuracy. However, being a recursive algorithm, split tracks cannot be reconnected. The second method showed how to use the estimated number of people in the scene for constraining the number of trajectories. Implemented with an offline tracker the tracking performance was improved.

The last part of this thesis presented three papers in the context of Smart Cities. Here it was demonstrated how thermal imaging can be used for robust performance in a number of real-time applications.

During the thesis, all methods have been tested on real world data and shown good and robust results over very long test periods. The detection and tracking algorithms have shown fast real-time performance and the future potential for use in real world applications is high.

14.1 Outlook and Perspectives

This thesis has just revealed solutions for a few of the infinite number of applications within analysis of humans. Research on analysis of people occupies a major part of the program in every conference on computer vision and will probably continue doing so for many years. When trying to mimic the abilities of the human vision system, we have only reached the most basic layers yet.

In this thesis methods for analysing humans in thermal images has been developed and applied. Choosing to work with thermal imaging has a clear advantage in real applications, as the privacy issues can be discarded. With thermal cameras, the use is no longer restricted to private ground, or to short sequences with staged activities. Without violating the privacy rights we have demonstrated applications, from sports arenas to city streets and plazas, with regular users and pedestrians captured.

Designing algorithms specifically for thermal imaging also implies that there is no standard test datasets with annotations available. That makes it harder to directly compare the results to other work. We have published the dataset used in chapter 5 and plan to publish the dataset used in chapter 8 and 9 with annotated tracking data. Over time, this might contribute to a larger thermal dataset used for state-of-the-art evaluation.

In this thesis we have focused on analysis of sports players, from the perspective of the use of sports arenas. Obvious potential is seen in applying some of the methods for other sports facilities, such as outdoor fields and trails. Focusing more on the players, the potential in sports analytics is also endless. Within the entertainment and broadcasting industry huge amounts of money are being invested to be able to provide unique and vivid experiences for the audience. For post-game analysis within the teams, coaches and athletes also need detailed evidence on their performance. As the research area develops, we expect to see more commercialised products within this field in the future.

SUMMARY

Measuring and mapping human activities are essential steps towards constructing an intelligent and efficient society. Using thermal imaging, the privacy issues often related to surveillance can be eliminated and public acceptance of such systems is easier to obtain. The main focus of this thesis is automatic analysis of the use of sports arenas. This work is organised under three themes: Occupancy analysis, Activity recognition and Tracking. Finally, the thesis demonstrates how thermal imaging can also be applied efficiently for analysing humans in the Smart City. This thesis starts by introducing the technology of the sensor and the different application areas. Two new methods for counting people are presented, as well as two methods for sports type recognition. Tracking of sports players is an important task in many applications, from recognition of activities to evaluation of performance. This thesis presents a real-time tracking algorithm based on Kalman filtering and it suggests two methods for improving global offline tracking. At the end of this thesis five different applications of thermal imaging in the Smart City are presented. Methods for counting and tracking pedestrians are presented and applied, as well as a method for detecting potential near-collisions between cars and cyclists in large urban intersections.

ISSN: 2246-1248
ISBN: 978-87-7112-194-0

AALBORG **UNIVERSITY PRESS**